DOCUMENT RESUME

ED 362 063 FL 021 554

AUTHOR Laurier, Michel
TITLE L'informatisation d'un test de classement en langue seconde (The Computerization of a Second Language Placement Test).
INSTITUTION International Center for Research on Language Planning, Quebec (Quebec).
REPORT NO ISBN-2-89219-234-X
PUB DATE 93
NOTE 275p.
PUB TYPE Reports - Descriptive (141) -- Reports - Research/Technical (143)
LANGUAGE French

EDRS PRICE MF01/PC11 Plus Postage.
DESCRIPTORS Comparative Analysis; *Computer Assisted Testing; *French; Higher Education; *Language Proficiency; Language Research; *Language Tests; Second Language Learning; *Student Placement; Test Construction; *Testing; Testing Problems

ABSTRACT

The purpose of this research was to compare two different forms of a placement test in French as a second language at the post-secondary level, a conventional test and a computerized test. First, expectations regarding a placement test as a measure of general proficiency are discussed and principles of adaptive testing that have been used to design the computerized test are presented. Then, the development of the two forms is explained and their psychometric properties are described. Finally, a theoretical and experimental comparison is made between the two forms. It appears that the computerized test is more accurate because of its focus and it does not give rise to any negative reactions from the students. (Contains over 300 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 362 063

dral icrlp
CENTRE INTERNATIONAL DE RECHERCHE EN AMÉNAGEMENT LINGUISTIQUE
INTERNATIONAL CENTER FOR RESEARCH ON LANGUAGE PLANNING

L'informatisation d'un test de classement en langue seconde

Michel Laurier

U S DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it
Minor changes have been made to improve reproduction quality
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Jean-Denis Gendron
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

FL021554

Publication B-190

FACULTÉ DES LETTRES

UNIVERSITÉ LAVAL

1993

L'informatisation d'un test de classement en langue seconde

Michel Laurier

B-190

1993

CENTRE INTERNATIONAL DE RECHERCHE EN AMÉNAGEMENT LINGUISTIQUE
INTERNATIONAL CENTER FOR RESEARCH ON LANGUAGE PLANNING
QUÉBEC

Données de catalogage avant publication (Canada)

Laurier. Michel (Michel D.)

L'informatisation d'un test de classement en langue seconde

(Publication B ; 190)
Comprend un résumé en anglais.
Comprend des réf. bibliogr.

ISBN 2-89219-234-X

1. Français (Langue) - Tests d'aptitude - Informatique. 2. Français (Langue) - Étude et enseignement - Évaluation. 3. Français (Langue) - Étude et enseignement - Allophones. 4. LISREL (Logiciel). I. Centre international de recherche en aménagement linguistique. II. Titre. III. Collection: Publication B (Centre international de recherche en aménagement linguistique) ; 190.

PC2066.L38 1993 448'.0076 C93-097089-6

Le Centre international de recherche en aménagement linguistique est un organisme de recherche universitaire qui a reçu une contribution du Secrétariat d'État du Canada pour cette publication.

The International Center for Research on Language Planning is a university research institution which received a supporting grant from the Secretary of State of Canada for this publication.

© CENTRE INTERNATIONAL DE RECHERCHE EN AMÉNAGEMENT LINGUISTIQUE
Tous droits réservés. Imprimé au Canada.
Dépôt légal (Québec) – 3e trimestre 1993
ISBN: 2-89219-234-X

À la mémoire du regretté

Michael Canale

dont les idées et les conseils ont

tracé la voie de cette recherche

REMERCIEMENTS

Je tiens à remercier sincèrement les membres de mon comité de thèse pour leur appui, leurs conseils et leur patience: le regretté Michael Canale, Les Mclean, Sharon Lapkin et Stacy Churchill. Je veux aussi remercier mon épouse Cécile pour sa compréhension et sa collaboration lors de la révision des textes. Je suis également reconnaissant à tous les étudiants qui ont accepté de faire les versions expérimentales du test ainsi qu'au responsables des programmes qui ont admirablement collaboré. Enfin, je dois souligner la participation financière de l'Université Carleton grâce à laquelle nous avons pu imprimer les versions finales.

RÉSUMÉ

Cette recherche avait pour but de comparer deux formes d'un instrument de classement en français langue seconde, au niveau post-secondaire: un test conventionnel et un test informatisé. On précise d'abord ce qu'on doit attendre d'un test de classement comme mesure de la maîtrise générale et on expose les principes du testing adaptatif qui ont servi à construire le test informatisé. Ensuite, on explique comment les deux formes ont été mises au point et on décrit leurs propriétés psychométriques. Enfin, on établit une comparaison théorique et expérimentale entre ces deux types de tests. Il ressort que le test informatisée s'avère plus court parce que mieux ciblé et qu'il ne suscite pas de réaction négative de la part des étudiants.

ABSTRACT

The purpose of this research was to compare two differents forms of a placement test in French as a second language, at the post-secondary level: a conventional test and a computerized test. First, expectations regarding a placement test as a measure of general proficiciency are discussed and principles of adaptive testing which have used to design the computerized test are presented. Then, the development of the two forms is explained and their psychometric properties are described. Finally, a theoretical and experimental comparison is made between the two forms. It appears that the computerized test is more accurate because of its focus and it does not give rise to any negative reactions from the students.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

INTRODUCTION

Au début des années '80, presque tous les intervenants dans le monde de l'éducation imaginaient, certains avec enthousiasme, d'autres avec appréhension, la place qu'occuperait l'ordinateur dans la salle de classe à la fin de la décennie. Dans le cadre d'une vaste enquête menée dans divers départements de langues modernes d'universités et de collèges américains, il y a maintenant plus de dix ans, Olsen (1980) indiquait que pas plus de 10% d'entre eux avaient recours à l'ordinateur. On pourrait penser qu'aujourd'hui la situation est radicalement différente. Or, force est de constater que la proportion n'a guère changé (Labelle 1986, Ng et Olivier 1987) et qu'en ce qui concerne l'enseignement de la langue seconde[1], on est encore loin du jour où la machine se substituera au professeur de langue ou s'intégrera simplement à l'environnement pédagogique comme le suggérait Amarel (1983). Les applications qu'on entrevoyait dans la classe de langue (Trush et Trush 1984, Holmes et Kidd 1982) sont encore l'initiative de praticiens marginaux oeuvrant dans des établissements plus fortunés. Ailleurs, les applications pédagogiques (A.P.O.) se bornent le plus souvent à des utilisations relativement simples des systèmes de traitement de texte (Freeman 1988).

Technologie encore trop rudimentaire? Financement inadéquat? Didacticiels peu intéressants? Méfiance générale? On peut invoquer plusieurs raisons sans toutefois pouvoir exclure définitivement l'intégration future de l'ordinateur à la classe de langue. La

[1] Comme notre recherche s'est déroulée dans le contexte de l'enseignement du français au Canada, nous parlerons dans la suite de l'exposé de langue «seconde» (F.L.S.). Il est évident que plusieurs de nos remarques et de nos conclusions s'appliquent tout autant à la situation de l'enseignement du français comme langue étrangère.

popularité des systèmes de traitement de texte est d'ailleurs le symptôme d'un malaise de la didactique des langues face à l'emploi de l'ordinateur. Comme l'indique Raschio (1986), il faut tout d'abord que la profession exprime explicitement ses attentes quant à la nature des didacticiels qu'on souhaiterait voir dans les classes de langue. À cet effet, Clark (1988) dresse une liste de priorités pour les prochaines années.

Comme spécialiste en évaluation de la langue seconde, Clark fait figurer au nombre de ces priorités, la mise au point de procédures de testing faisant appel à l'ordinateur. Dans un article publié quelques années plus tôt (Clark 1983), il signalait d'ailleurs que l'exploitation de l'ordinateur constituait, avec l'évaluation directe et authentique, une des voies d'avenir dans le domaine du testing en langue seconde. Il ne fait pas de doute que les traditionnels tests «papier-crayon» demeureront, mais l'ordinateur peut présenter une alternative intéressante dans certains cas. Ainsi que le font remarquer Higgins et Johns (1984:97):

> It is unlikely that the computer will, for the foreseeable future, replace paper and pencil as the direct medium for the student to use in mass test. (...) For the individual student the computer is, of course, an ideal medium for self-testing, providing diagnostic information that may be rough-and-ready or fairly sophisticated.

On peut imaginer certains des avantages d'une administration informatisée par rapport à une administration plus traditionnelle: confidentialité du test, correction automatique et immédiate, individualisation de la procédure, administration à un seul sujet... Mais il faut se demander si ces avantages sont bien réels dans le cas d'un test de langue, s'ils ne sont pas obnubilés par des contraintes trop nombreuses ou trop sérieuses. D'où la question fondamentale à laquelle la présente recherche tentera de répondre: est-il possible d'utiliser de façon avantageuse le micro-ordinateur pour l'évaluation de la langue seconde? *A priori*, nous sommes tentés de répondre affirmativement bien qu'une telle réponse contienne une large part d'incertitude et d'hésitation. Encore faut-il préciser à quelles fins le

test sera utilisé et dans quelles conditions. Il faut se demander si l'emploi d'une technologie ne se fait pas au prix de concessions quant à la pertinence de l'instrument ou n'a pas d'implications indésirables au plan psychométrique.

Dans le contexte actuel des recherches dans ce domaine, tout est à faire pour qui veut observer les avantages et inconvénients d'un test informatisé en français langue seconde. Comme il n'existait pas, au moment où nous avons entamé le présent projet, de test standardisé offrant à la fois une version «papier-crayon» et une version informatisée et pouvant être utilisé auprès de notre population cible, il a fallu développer un instrument de mesure. C'est pourquoi, le présent document adopte une organisation tripartite, qui respecte la démarche que nous avons suivie au cours de la recherche:

- Un aspect théorique définissant les objectifs et l'organisation du test.
- Un compte rendu du processus d'élaboration des deux versions du test.
- Une comparaison entre les administrations des deux versions du test.

L'aspect théorique sera couvert par les deux premiers chapitres. Dans le premier, on situera notre recherche dans le cadre de la problématique actuelle de l'évaluation de la langue seconde. On en viendra alors à préciser de quel type de test il s'agit, compte tenu des usages anticipés et de la population visée. Le second chapitre traitera des modèles psychométriques qui sous-tendent l'élaboration d'un test adaptatif. Nous justifierons alors nos options théoriques en ce qui concerne la version informatisée.

On trouvera dans les troisième et quatrième chapitres le compte rendu de l'élaboration du test depuis la version pré-expérimentale jusqu'aux versions finales, «papier-crayon» et informatisée. Il s'agira de retracer les étapes, longues mais nécessaires, de la mise en place des outils permettant l'expéri-

mentation finale. Nous discuterons des caractéristiques psychométriques des instruments et de la programmation de la version sur ordinateur.

Les deux derniers chapitres correspondront à la troisième partie de notre démarche. Ces chapitres représentent en quelque sorte le but ultime de la recherche. Il s'agira de comparer l'administration de la version informatisée par rapport à celle de la version traditionnelle aux plans psychométrique, psychologique et administratif. Dans le cinquième chapitre, nous comparerons les deux versions du point de vue théorique. Dans le sixième chapitre, nous considérerons plutôt des données expérimentales.

C'est ainsi que la perspective adoptée place la présente recherche au confluent de trois disciplines principales: la didactique des langues, la technologie éducative et la docimologie. Toutefois, c'est plus particulièrement à la première, la didactique des langues, que nous espérons apporter une contribution. En d'autres termes, nous comptons mettre au service de la didactique des langues, des innovations dans les domaines de la technologie éducative et de la docimologie.

①

L'ÉVALUATION DE LA MAÎTRISE EN LANGUE SECONDE

1.1 Les débats actuels

On ne s'étonnera pas que les développements relatifs à l'évaluation de la langue seconde suivent de près l'évolution des approches et des méthodologies. Il est d'ailleurs souhaitable qu'il en soit ainsi puisqu'il est essentiel que les pratiques évaluatives s'accordent avec les objectifs pédagogiques. Néanmoins, rien n'indique *a priori* que l'évaluation doive être à la remorque des autres domaines de la didactique des langues. Pas plus, pour reprendre la métaphore de Roe (1981), que le testing ne doit se comporter comme le «coucou dans le nid» c'est-à-dire déloger les préoccupations fondamentales de la salle de classe en imposant ses propres critères. On doit se réjouir que plusieurs des débats que connaît présentement le domaine de l'évaluation trouvent leur source dans les grandes controverses qui animent la didactique. Cependant, il est évident que certaines questions ont pris, sous l'angle de l'évaluation, une importance et une direction originales. Élaborer un test implique qu'on prenne position par rapport à ces débats. C'est pourquoi, dans les pages qui suivent, nous allons faire le point sur les courants actuels dans l'évaluation de la langue seconde.

1.1.1 *La réalisation des tâches langagières*

Comme le souhaitaient Chastain (1977) ou Fishman et Cooper (1978), les principes de l'enseignement dit «communicatif» n'ont pas tardé à s'imposer au niveau de l'évaluation. Dans cette

perspective, Johnson (1979) insiste sur la nécessité de recréer des interactions naturelles tant dans les activités d'enseignement que dans les activités d'évaluation. Il dégage ainsi trois principes fondamentaux:

- Les activités doivent impliquer la réalisation de tâches.
- Tout échange suppose la transmission d'information.
- L'étudiant[1] intervient dans la sélection des activités.

Le premier principe établissait d'une façon incontestable la supériorité du testing direct qui vise à reproduire une situation authentique, par rapport au testing indirect qui sert davantage d'indication (Clark 1975). Jones (1985:3) décrit ainsi l'évaluation directe de la performance: *Its purpose is usually to assess the ability of an examinee in relation to some kind of job-related task. (...) The overall criterion is the succesful completion of a task in which the use of language is essential.* À titre d'exemple de test axé sur la réalisation de tâches, mentionnons l'épreuve orale que doivent subir les assistants d'enseignement venus de l'étranger pour étudier à l'Université de Californie (UCLA) (Bailey 1985, Hinofotis *et al.* 1981) ou les tests mis au point sous la supervision du *English Language Testing Service (ELTS)* conformément aux exigences et aux contenus de programmes d'études particuliers (Carroll 1985, Ingram 1990). Au Canada, certains tests de français langue seconde, conçus dans le cadre de législations linguistiques, visent à évaluer la capacité à communiquer dans la langue seconde en milieu de travail et s'inscrivent dans ce courant (Gareau 1981, Monfils 1982). On note également une tendance à évaluer à partir de tâches communicatives, dans le milieu scolaire (Girard *et al.* 1984, Lussier-Charles et Danan 1983, Jasmin-Demers 1983, Lapkin 1985).

Dans la perspective du testing direct, il devient important d'intégrer la réalisation de tâches aux tests de langue tout en

[1] Dans ce document, nous parlerons d'«étudiant», de «sujet», de «professeur»... Le masculin est utilisé sans discrimination, dans le but d'alléger le texte et de respecter la norme du français.

gardant à l'esprit que des étudiants pourront utiliser des moyens différents pour accomplir des tâches similaires (DesBrisay 1981, Lantoff et Frawley 1988). Les activités d'évaluation ressemblent de plus en plus aux activités de classe lesquelles se modèlent sur des situations susceptibles d'être vécues par l'étudiant. Ainsi que l'explique Potts (1985) il s'établit, dans le cadre d'une approche communicative basée sur les tâches, une dialectique enseignement/évaluation.

Par ailleurs, dans une évaluation directe, la validité du construit est la seule exigence psychométrique qui tienne. Comparant les tests communicatifs aux tests classiques Harrison (1983a:80) écrit: *Good communicative tests (like other tests) are straightforward tasks for the student to do, not Machiavellian sorting systems.* Dans la même veine, il en arrive à mettre en doute le fait que le testing puisse prétendre à l'exactitude et à l'objectivité. De fait, il est certain que la validité de ce type de test ne peut pas être définie par les indices psychométriques classiques (Low 1985).

Outre les problèmes psychométriques inhérents aux tests basés sur la réalisation de tâches spécifiques, Genesee (1982,1984) signale trois problèmes auxquels sont confrontés les utilisateurs.

- Comment s'assurer, dans la pratique, que la situation de test atteindra le degré d'authenticité souhaité?
- Peut-on généraliser l'évaluation à d'autres situations auxquelles l'étudiant pourrait faire face?
- Quels critères adopter lors de la définition des niveaux de performance requis?

Par ailleurs, il faut signaler que l'élaboration de tests spécifiques dans le cadre d'une organisation où les intervenants sont appelés à exercer une grande variété de tâches peut devenir un problème insurmontable (Ricciardi 1981, Wesche 1987). Shohamy et Reves (1985) font remarquer qu'il faut viser l'authenticité mais que les contraintes qu'implique cette orientation justifient certains compromis.

Cette recherche de l'authenticité à travers la réalisation de tâches vraisemblables et susceptibles de se produire n'est pas une absolue nécessité lors d'une évaluation formative dans laquelle on cherche à diagnostiquer des problèmes particuliers ou certains acquis. Elle devient une exigence fondamentale lors d'une évaluation sommative dans laquelle «seul compte le résultat» (Mothe 1985:60)

1.1.2 *L'intégration des éléments*

Traçant l'historique de la notion d'évaluation de la langue seconde, Splosky (1976) distingue trois périodes principales:

- La période pré-linguistique: ce type d'évaluation cadrait avec une approche qui mettait l'accent sur un apprentissage cognitif et sur l'analyse contrastive.
- La période psychométrique-structurale: issue des principes de la linguistique structurale et des recherches psychométriques, ce mouvement privilégiait la notion d'unité discrète.
- La période sociolinguistique-intégrative: dans la lancée de l'approche communicative, cette tendance s'oriente vers l'évaluation de la maîtrise de la langue, en situation.

Une des caractéristiques de la dernière période est le fait qu'on rejette le postulat voulant que la compétence se mesure en faisant la somme des éléments linguistiques acquis (Groot 1975). On estime que l'utilisation effective de la langue suppose la capacité de produire ou de comprendre les éléments linguistiques en interrelation les uns avec les autres dans un contexte particulier.

Bien que promoteur d'une vision fondée sur la multiplicité des facettes de la performance linguistique, Carroll (1965) reconnait le premier la valeur des tests intégratifs pour mesurer les interactions entre les facettes. Quelques années plus tard, Carroll (1982) en viendra à considérer le test intégratif non seulement comme un

moyen de ne plus isoler les points de langue mais aussi comme le moyen de transgresser les distinctions entre les quatre savoir-faire (Leblanc 1983); la globalité de la performance se trouve ainsi préservée.

Dans un ordre d'esprit tout à fait différent, on verra dans les tests intégratifs, l'occasion de mesurer l'essentiel de la compétence du sujet. Oller (1978:46) décrit cette habileté langagière fondamentale qu'il nomme «grammaire de l'expectative» comme *a device that generates and confirms hypotheses*. Oller (1972) ajoute que la «grammaire de l'expectative» ne s'évalue pas nécessairement à l'aide de tests directs. En d'autres termes, un test intégratif n'est pas forcément communicatif (Savignon 1983:chap 6, Spolsky 1985). En effet, des mesures indirectes telles que la dictée, traditionnelle (Oller et Streiff 1975, Cziko 1982) ou avec brouillage (Gradman et Spolsky 1975), sont des indications fiables du degré d'acquisition de la capacité de générer et de confirmer des hypothèses. Oller (1973) démontre que le test de closure constitue également une mesure intégrative, indirecte mais efficace. À l'appui de la position défendue par Oller, Palmer (1983) distingue entre un contrôle «compartimenté» (par éléments discrets) et un contrôle «intégré»; il conclut en la supériorité d'un programme qui développe le contrôle intégré.

Il est certain que de nombreux procédés discursifs, des relations grammaticales inter-phrastiques, des références textuelles ou situationelles ne peuvent s'évaluer sans le recours à un test intégratif. Cazabon (1984) signale même que le test à éléments discrets tend à figer l'usage alors que le test intégratif, doté de contexte, accorde une place à la variation linguistique.

Cette distinction entre test intégratif et test à éléments discrets ne fait toutefois pas l'unanimité. Ingram (1978) ne trouve pas de différence significative entre les corrélations de tests intégratifs et celles de tests à éléments discrets. Farhady (1983d) arrive à des observations semblables de sorte qu'il conteste l'existence de la distinction d'un point de vue statistique.

1.1.3 *L'hypothèse du trait unitaire*

1.1.3.1 *Les modèles factoriels*

Briere (1969) rapporte qu'au cours d'un congrès tenu en 1968, Spolsky et Upshur signalaient tous deux la possibilité de mesurer un facteur sous-jacent dans le cadre d'une théorie fondée sur la communication. Dans cette optique, communiquer ferait intervenir une habileté langagière fondamentale que partagent les locuteurs et sur laquelle s'appuient les manifestations linguistiques plus spécifiques comme le mode (écrit ou parlé), le registre, les fonctions... Par ailleurs, dans le cadre d'une approche basée sur la réalisation de tâches, Carroll (1980) établit des corrélations entre différents tests de performance et observe la présence d'un facteur général qui rendrait compte de 58% de la variance.

Pour Oller, ce facteur général est en fait la «grammaire de l'expectative» que tentent de mesurer les tests intégratifs.

> Within the context of expectancy grammar as models of underlying competence, a valid language test can be defined as one that activates the expectancy grammar that the learner has internalized. The extent to which the learner's grammar is able to synthesize and analyse meaningful sequences of elements of language is an indication of his proficiency or competence in the language. (Oller 1978:52)

Dans cette perspective, Oller (1981) en vient à penser que la valeur d'un test de performance général tiendrait en grande partie à la manière dont l'instrument mesure cette compétence sous-jacente.

La question du facteur général devient une préoccupation centrale avec la parution de deux ouvrages importants (Oller 1979, Oller et Perkins 1980) où Oller défend l'hypothèse du trait unitaire. Oller évoque d'ailleurs la possibilité que ce trait unitaire soit associé au facteur «g» qui a suscité, il y a quelques années, de vives controverses dans le domaine de la psychologie. On parle alors d'un facteur général qui rendrait compte de 65% de la variance entre des

mesures aussi diverses que le *Test of English as a Foreign Language (TOEFL)*, le test du *Center of English as a Second Language* de l'Université d'Illinois *(CELT)* ou l'entrevue du *Foreign Service Institute (FSI)* (Oller et Hinofotis 1980). Utilisant une procédure d'analyse factorielle similaire à celle de Oller pour comparer 22 mesures différentes, Scholz *et al.* (1980) arrivent à des résultats du même ordre.

L'hypothèse du trait unitaire est cependant contestée. Carroll (1983, 1987) rappelle que le facteur général est souvent un artéfact de la méthode d'analyse. Il soumet les données de Scholz *et al.* à une analyse factorielle différente. Plutôt que de recourir à une procédure de recherche des composantes principales, il utilise une procédure de décomposition en facteurs. Sans infirmer l'hypothèse du trait unitaire, il réduit l'importance du premier facteur et dégage un deuxième et un troisième facteur, respectivement le mode écrit et le mode oral. Farhady (1983b) remet en question la procédure de Oller sur le plan méthodologique en soutenant qu'il faut non seulement utiliser la décomposition en facteurs, mais qu'il faut aussi effectuer la rotation des facteurs. En procédant ainsi, il conclut que l'hypothèse du facteur unitaire devient beaucoup moins plausible.

À l'instar de Carroll et Farhady, Woods (1983) met en doute la pertinence de la procédure de recherche des composantes principales en soulignant qu'elle tend à favoriser l'émergence d'un facteur principal. Par contre, la décomposition en facteurs a une valeur plutôt confirmatoire en ce qu'elle suppose que l'utilisateur fournisse un modèle à vérifier. De plus, elle mène souvent à des résultats difficiles à interpréter. Vollmer (1981) rappelle aussi, fort justement, que les données de base sont toujours des scores obtenus à un test et que rien n'indique que les conclusions puissent s'appliquer à l'usage réel de la langue à des fins communicatives. Il ne faut d'ailleurs pas s'étonner de la persistence d'un facteur général puisque tout test implique une composante verbale plus ou moins importante (Streiff 1983). Plus encore, beaucoup de tests qui comportent une composante verbale mesurent un facteur commun qu'on a souvent associé à l'intelligence (Jensen 1980:chap 5) et qui, de ce fait, a peu à voir avec la compétence en langue seconde. Les

critiques d'ordre méthodologique vont jusqu'à mettre en cause le bien-fondé de l'analyse factorielle. En l'absence de procédure statistique appropriée, Vollmer (1983) refuse d'admettre l'hypothèse du trait unitaire. Il ajoute que même lorsqu'on l'a observé, ce facteur demeure un concept inopérant puisqu'il s'agit d'une abstraction qu'on ne peut définir.

À la suite des nombreuses interrogations soulevées par l'idée d'un facteur général, Oller en vient à proposer que ce facteur général puisse lui-même être composé de plusieurs habiletés. Toutefois, s'il tempère la théorie du trait unitaire, Oller (1983) n'en maintient pas moins le principe. Cette redéfinition corrobore les vues de Upshur et Homburg (1983) pour qui le résultat d'un test à plusieurs parties, se compose de l'apport d'un facteur unique et de facteurs spécifiques pour chaque sous-test.

Dans une recherche plus récente, Davidson (1988) estime que et les conclusions de Oller étaient exactes. Il ajoute cependant que le modèle théorique ne doit pas nécessairement être validé par une analyse factorielle dont on reconnaît maintenant les insuffisances. Il introduit une distinction entre, d'une part, la notion de «facteur» (ou de «dimension»), une création statistique, et d'autre part, la notion de «facette», un concept théorique. Ainsi, le fait qu'on isole un facteur général dans les tests de langue, n'empêche pas qu'on puisse reconnaître une multiplicité de facettes. Comme Hulstijn (1985) il reconnaît les limites des recherches empiriques en indiquant que les modèles théoriques demandent à être validés par d'autres moyens que l'analyse factorielle.

1.1.3.2. *Les modèles théoriques*

Les recherches de Bachman et Palmer témoignent d'une réflection pour mettre en place un modèle théorique acceptable et opérationnel. Leurs premières études menées en utilisant une analyse factorielle reconnaissaient l'existence d'un facteur général mais aussi l'importance de facteurs spécifiques: l'expression orale par rapport à la lecture (Bachman et Palmer 1981) ou un facteur

grammatical/pragmatique par rapport à un facteur sociolinguistique (Bachman et Palmer 1982). Ils ont par la suite construit un modèle théorique reposant sur l'idée que tout test de langue mesure un facteur général et divers facteurs spécifiques reliés à la situation de communication ou au test lui-même (Bachman et Palmer 1984).

En raffinant le modèle, Bachman (1990) propose une compétence langagière composée de deux traits:

- une compétence organisationnelle, elle-même divisible en une compétence grammaticale et une compétence discursive; on estime que la compétence organisationnelle contribue à 30% du résultat d'un test de langue.
- une compétence pragmatique, elle-même divisible en une compétence illocutionnaire et une compétence sociolinguistique; le poids relatif de la compétence pragmatique serait de 25%. Par ailleurs, le résultat d'un test manifeste aussi la présence de diverses habiletés langagières parmi lesquelles figurent les stratégies qu'utilise le sujet; ces habiletés représentent environ 30% du résultat. Enfin tout résultat comprend des effets reliés au test lui-même: il s'agit d'effet de méthode et d'effets aléatoires.

Ce modèle théorique n'est pas sans rappeler le modèle désormais classique proposé par Canale et Swain (1980). Ce modèle postule dans sa version finale une compétence communicative elle-même formée de quatre compétences plus spécifiques:

- la compétence grammaticale,
- la compétence sociolinguistique,
- la compétence de discours,
- la compétence stratégique.

L'importance accordée à chacune des composantes de la compétence communicative varie selon les activités pédagogiques proposées (Canale 1983) ou selon les formes d'évaluation mises en oeuvre (Canale 1981a). Le modèle de Canale et Swain demeure un outil d'analyse et de développement précieux; plus qu'à des données statistiques, sa validité tient au fait qu'il reflète les contributions de la linguistique, de la sociolinguistique et de la psycholinguistique de même que les priorités actuelles dans l'enseignement des langues.

Courchêne et de Bagherra (1981) formulent deux objections majeures face au modèle théorique de Canale et Swain. D'une part, ils s'interrogent sur le statut de la compétence stratégique particulièrement dans une perspective évaluative. De fait, il vaut peut-être mieux s'inspirer du modèle de Bachman et Palmer et ranger les stratégies du sujet parmi les habiletés. D'autre part, Courchêne et de Bagherra signalent que le modèle semble négliger les interrelations qui lient les diverses compétences. La remarque est d'autant plus pertinente qu'on ne peut ignorer les conclusions des analyses factorielles. Si on veut maintenir la divisibilité de la compétence langagière, on doit alors admettre que les composantes de cette compétence soient fortement corrélées. Accepter l'idée d'un tel réseau d'interconnections entre les composantes, c'est reconnaître la diversité déroutante des interactions langagières et la complexité indéniable de l'acquisition d'une langue. Faisant le point dans le débat, et tentant de réconcilier l'approche factorielle et l'approche théorique, Carroll (1983:94,103) écrit:

> The general proficiency factor reflects overall degree of advancement in different language skills - as a function of the way the language is taught, the attention and effort the learner devotes to the study of the language, and possibly (or probably) the rate at which the learner is able to absorb and master what is being taught. (...) Such evidence as is available suggests that specialized verbal skills are learned, and correlations among these skills tend to index the extent to which they tend to be learned together.

1.1.4 *Le concept de maîtrise*

Au cours des dernières années, on a vu se développer aux États-Unis, un mouvement autour de la notion de maîtrise[2]. Le mouvement a pris d'autant plus d'ampleur qu'il a l'appui du puissant *American Council for Teaching Foreign Languages (ACTFL)*. Les principes de base de ce mouvement (Omaggio 1983a) sont fortement inspirés par le testing. Le concept de maîtrise se présente dans une certaine mesure comme la synthèse des trois débats que nous venons de décrire.

Certes, le concept privilégie le testing direct. À mesure que se développe la maîtrise, les tests doivent devenir de plus en plus intégratifs. À cet égard, Omaggio (1983b) recommande de contextualiser les tests de langue qu'utilisent les enseignants dans les classes de langue. Toutefois, on élargit la notion de tâche pour inclure des tâches plus linguistiques que communicatives. Les opposants de la notion de maîtrise ne manquent pas d'y voir une façon de réintégrer l'évaluation par éléments discrets (Savignon 1985, Bachman et Savignon 1986). Par ailleurs, face au problème de généralisation que pose la réalisation de tâches spécifiques, la popularité de la notion de maîtrise témoigne d'un désir de dépasser la situation à partir de laquelle s'établit l'évaluation directe. Lowe (1980) ne manque pas de souligner que la maîtrise implique la capacité de généraliser vers plusieurs situations. En d'autres termes, la performance devient un indicateur de la maîtrise.

Rivera (1982) s'interroge sur cette notion un peu floue qui ne correspond ni à la compétence, ni à la performance. Parle-t-on de la connaissance du code ou de l'usage approprié de ce code? Ingram (1985:223) fournit un élément de réponse:

> One can distinguish the underlying general proficiency a learner has in a particular macroskill from the learner's

[2] À défaut de terme plus satisfaisant nous parlons de «maîtrise», pour nous référer au sens qu'a pris le terme *proficiency* dans le cadre de ce mouvement. En dehors de celui-ci, la notion se confond parfois avec la notion de performance.

> ability to carry out an absolutely specified task in a specified situation. (...) General proficiency would seem to entail the ability to use commonly occuring features (e.g. phonology, syntax, lexis, discourse, functions etc.) and would seem to underlie the learner's register flexibility or his ability to cope with new situations and to select within his language repertoire to modulate his language according to situational need.

Ainsi, il semble que la maîtrise soit sous-jacente à toute utilisation de la langue dans une situation de communication. Toutefois, elle se distingue du facteur unitaire de Oller qui se situait nettement au niveau de la compétence. Situation problématique puisque, comme le souligne Vollmer (1981), la notion de maîtrise se situant entre les notions de compétence et de performance, elle devient de ce fait non-observable et, partant, peut-être inutile.

Lowe (1985) rappelle que la notion de maîtrise est issue des travaux sur l'évaluation de la langue seconde, menés par l'*Inter-Agency Language Roundtable (ILR)*. De fait, on vise une interdépendance de l'enseignement et de l'évaluation: *We test what we teach, we also teach what we test* (Magnan 1985:143). Plutôt que de définir la maîtrise, Lowe en donne trois caractéristiques, mettant ainsi en évidence le caractère empirique de la notion:

- plutôt qu'un système à éléments discrets, elle est globale;
- elle implique plusieurs facteurs implicites;
- elle dépasse le simple rendement.

Bien que soulevant de sérieux problèmes théoriques, la notion de maîtrise n'en est pas moins intéressante. D'une part, elle se distingue de la compétence que les linguistes et les sociolinguistes définissent comme un ensemble de règles. Or, non seulement la compétence est difficile à mesurer mais on sait que la connaissance de règles ne saurait être une fin en elle-même. D'autre part, la maîtrise se distingue de la performance qui, dans la perspective de Chomsky (1965:4), se veut la réalisation concrète de l'application des règles dans une situation particulière. L'évaluation de la perfor-

mance risque donc d'aboutir à des résultats qui ne seront pas généralisables. La maîtrise se définit comme la capacité de mettre en oeuvre une compétence en vue d'une performance. La maîtrise apparaît donc comme le moyen terme, qui devient l'objectif à enseigner et l'objet à évaluer.

1.2 Un test de classement

1.2.1 *Les objectifs du test*

Avec la popularité croissante des cours de langue dans les établissements post-secondaires, il devient essentiel de compter sur des tests de classement adéquats. En effet, il est fréquent que des établissements se retrouvent devant un grand nombre d'étudiants qui doivent être classés tant bien que mal, en quelques heures, et placés dans le niveau qui convient à chacun[3].

Or, quand on fait l'inventaire des tests disponibles au Canada pour le français comme langue seconde ou étrangère, on s'aperçoit qu'il s'y trouve bien peu de tests de classement (Savard 1969 et, pour les tests plus récents, Lapkin *et al.* 1984). L'annuaire du *Buros Institute of Mental Measurement* (Mitchell 1983) signale bien quelques tests de classement de niveau collégial, encore disponibles aux États-Unis — dont la batterie élaborée par le *Educational Testing Service (ETS)*. Pourtant, bien qu'on reconnaisse parmi les spécialistes de la didactique des langues, l'importance de l'adéquation entre l'instrument de classement et l'approche pédagogique (Dermer-Applebaum et Taborek 1986), on ne dispose que de tests désuets qui ne tiennent pas compte des derniers développements en didactique des langues.

Face à la pénurie de tests de classement en F.L.S., beaucoup d'établissements se voient contraints d'utiliser des tests qui ne cadrent plus avec les pratiques pédagogiques qui ont cours. Ces tests font généralement partie des instruments mis au point au

[3] Nous traduisons l'expression anglaise *placement test* par «test de classement». Les tests de classement servent à former des groupes homogènes et non pas à opérer une forme de sélection parmi un groupe d'étudiants.

cours de la période psychométrique-linguistique ou même pré-linguistique. Le test Laval (Gendron *et al.* 1971) et le test de la Commission des écoles catholiques de Montréal (Douesnard *et al.* 1972) par exemple, s'utilisent encore. Ces tests se concentrent sur une seule des quatre composantes de la compétence communicative telle que décrite par Canale et Swain à savoir la compétence linguistique. Par contre, on trouve pour l'instant, peu de tests qui mesurent de façon satisfaisante la maîtrise, c'est-à-dire la capacité d'utiliser la langue en situation, bien que les bases théoriques de tels tests soient assez clairement établies.

Diverses solutions au problème de classement ont été proposées: courtes entrevues (Ilyin 1970), recours à l'auto-évaluation (Painchaud et Leblanc 1984, Leblanc 1985), classement d'après le dossier scolaire, ou même suppression de tout processus de classement. Toutefois, il s'agit, dans plusieurs cas, de solutions imparfaites ou *ad hoc*, à défaut d'un mode d'évaluation plus convenable. Par «convenable», on entend d'abord un test qui soit valide c'est-à-dire qui soit cohérent avec le type d'enseignement qui prévaut habituellement, puis un test qui soit fiable afin d'éviter les changements de groupe trop nombreux et, finalement, un test qui soit commode de sorte qu'on puise classer rapidement, économiquement et simplement un grand nombre d'élèves. De plus, il y a fort à parier que dans les milieux de l'éducation, l'exigence de commodité a préséance sur celles de validité et de fiabilité.

Il faut aussi tenir compte de la diversité des programmes offerts par les établissements et de la mobilité de la population étudiante. Construire un test implique alors la recherche de la «mesure commune» pour reprendre l'expression de Clark (1980) duquel, du reste, nous nous inspirerons dans notre démarche. Il ne fait donc aucun doute que la création d'un test en F.L.S. répond à un besoin dans le domaine de l'enseignement aux jeunes adultes. Cela est d'autant plus vrai que ceux-ci, contrairement à leurs cadets, ont connu, quand on les accueille, des apprentissages fort différents les uns des autres, ce qui rend problématique l'estimation de leur niveau de compétence.

Par ailleurs, l'élaboration d'un test qui fasse appel aux ressources de la micro-informatique présente un intérêt certain dans le domaine de l'enseignement des langues secondes. À l'heure où les laboratoires de micro-ordinateurs sont en voie de remplacer les laboratoires de langues devenus apparemment désuets, à l'heure où l'équipement informatique pourrait faire de plus en plus partie de la panoplie des services pédagogiques et où chacun se demande si cette technologie est d'une quelconque utilité dans l'enseignement de la langue, une application comme la nôtre semble la bienvenue. Elle pourrait compléter la bibliothèque, encore bien limitée et de qualité très inégale, d'outils pédagogiques qui exploitent les possibilités des micro-ordinateurs.

La population étudiante que nous visons particulièrement est principalement constituée d'étudiants qui s'inscrivent dans des établissements post-secondaires, à des cours de langue, soit pour satisfaire les exigences d'un programme d'études, soit comme complément à leur formation personnelle ou professionnelle. Il s'agit d'élèves qui sont sur le point de terminer leurs études secondaires ou qui sont déjà engagés dans des études de niveau collégial ou universitaire. Toutefois, on peut facilement imaginer qu'un tel test de classement puisse aussi servir pour l'ensemble des élèves du niveau secondaire ou auprès de la population adulte en général.

Cependant, la standardisation du test et la comparaison entre les deux versions du test ont été menées auprès d'un échantillon relativement homogène. Il s'agissait d'étudiants inscrits dans des programmes intensifs offerts pendant l'été. La grande majorité étaient des boursiers du Secrétariat d'État. Chaque année des milliers de ces étudiants ont la chance, grâce à une bourse provenant du Secrétariat d'État, de vivre six semaines d'immersion totale, en dehors de leur milieu habituel (Keating 1989). Les établissements qui participent au programme de bourses doivent offrir une vingtaine d'heures de cours de langue par semaine. Ces cours visent le développement de la performance plutôt que la connaissance du code linguistique et mettent nettement l'accent sur les habiletés orales. Plusieurs activités para-scolaires telles que

des excursions, des spectacles et des ateliers se greffent au programme. Durant ces activités, on insiste sur l'usage exclusif de la langue cible.

Les récipiendaires des bourses pour l'apprentissage du F.L.S. représentent une population relativement homogène du point de vue sociologique. La très grande majorité sont âgés de 17 à 22 ans. Ils sont tous inscrits comme étudiants à temps plein dans des établissements de niveau secondaire ou post-secondaire canadiens où l'anglais est la langue d'enseignement. Ces boursiers choisissent de consacrer une partie de la période estivale à l'apprentissage du français. Lorsque nous visitions un établissement, nous avions l'habitude d'expérimenter le test auprès d'un échantillon représentatif ou auprès de l'ensemble des étudiants. Il n'y avait donc aucune raison de suspecter que l'échantillon puisse être faussé par la sur-représentation d'un domaine de spécialité par exemple (Alderson et Urquhart 1983 1985, Farhady 1983a) ou par des variations dans la langue de départ (Ramirez 1984). Par contre, on trouve dans la population étudiée de grandes variations quant au niveau de performance dans la langue cible. L'étape du classement s'avère donc déterminante pour le succès de tels programmes.

1.2.2 *Spécificité du test de classement*

Qu'attend-on d'un test de classement? Essentiellement qu'il départage, au début d'un programme, une masse d'étudiants en différents groupes afin que tous puissent bénéficier d'un enseignement approprié. Cette vocation du test de classement le distingue à plusieurs égards des autres types de test.

1.2.2.1 *La validité prédictive*

Fournir un programme approprié, c'est faire en sorte que l'apprenant soit intégré à un groupe où il pourra progresser le plus possible et ce, tout en suivant le rythme général du groupe. Comme le test de classement cherche à déterminer si un programme conviendra ou non, on ne peut juger de la valeur du classement

qu'après coup, en utilisant des données qui s'accumulent pendant le programme. Il peut s'agir du jugement de l'apprenant ou de son professeur après quelques heures d'enseignement; on peut considérer, des tests périodiques administrés en classe; on peut comparer le classement avec la note finale au cours. Dans tous les cas, le test de classement sert à prédire le comportement ou le degré de succès de l'apprenant. Le fait que le test de classement précède l'enseignement implique qu'on ne peut pas l'intégrer aisément au programme de cours selon les principes d'une évaluation naturelle telle que décrite par Canale (1984, 1985, 1988). Par conséquent, on doit obligatoirement prendre en considération la validité prédictive du test de classement.

Dans cette perspective, plusieurs types de renseignements permettent de prédire si le programme conviendra. Beaucoup d'étudiants s'attendent à être classés en fonction de leur nombre d'années d'étude dans la langue seconde. La note obtenue au dernier cours de langue peut également servir de base au classement. Currall et Kirk (1986) rapportent le cas d'un programme où l'élément qui prédisait le mieux les chances de succès à un cours de langue était les résultats scolaires obtenus dans les autres cours. On peut choisir de regrouper les apprenants en fonction de leurs besoins ou des intérêts qu'ils ont exprimés au début du cours. On peut même imaginer que le classement s'établisse à partir de données personnelles comme l'âge, la personnalité, la motivation...

On peut s'appuyer sur les aptitudes et les stratégies d'apprentissage. Par exemple, la division de la formation linguistique de la fonction publique du Canada (Monfils 1982) tient compte des résultats de tests d'aptitude pour l'apprentissage d'une langue seconde comme le fameux *Modern Language Aptitude Test (MLAT)*, mis au point par Carroll et Sapon (1959). Harris (1970) montre que la simple mesure de la mémoire à court terme peut être une donnée suffisante à des fins de classement. Chapelle et Roberts (1986) estiment que le succès à un cours de langue peut être relié à certaines caractéristiques cognitives. Il est donc possible de déterminer les stratégies et les styles d'apprentissage et de faire en sorte qu'on réunisse, par exemple, les apprenants qui privilégient une approche

analytique dans un groupe différent de ceux qui sont plus à l'aise avec une approche globale. On pourrait aussi déterminer la prédominance de certaines stratégies d'apprentissage (O'Malley *et al.* 1985) et en tenir compte lors du classement.

Le critère, ou l'ensemble de critères, qui sert à établir le classement peut donc varier. Dans le cadre de la théorie de la décision, Hills (1971:714) précise: *Thus the extent to which the test permit useful placement is a joint function of how well the test measures the relevant underlying trait and how markedly the validity of the trait differs for different treatments.* Il est permis de croire que la maîtrise porte une haute valeur prédictive. Tyler (1974) indique que ce n'est qu'à partir du milieu des années cinquante que le regroupement en fonction du niveau d'habileté est devenu une préoccupation générale dans le domaine de l'éducation; il met d'ailleurs en doute la supériorité de ce mode de classement.

Enfin, si on en est venu à identifier le test de classement à une mesure du niveau d'habileté, il n'en reste pas moins que cette évaluation peut prendre diverses formes. Pour un programme de compréhension auditive au niveau débutant, une mesure de la discrimination auditive peut être hautement prédictive alors que pour un programme comportant des objectifs structuraux, une mesure de la compétence grammaticale constituera une meilleure indication des résultats anticipés. Swain *et al.* (1974) font remarquer que dans certaines circonstances, la traduction et l'imitation peuvent être des activités qui réflètent bien la compétence en langue seconde. Par contre, en vue de l'atteinte d'objectifs communicatifs, il faudrait songer à une évaluation de la maîtrise générale.

1.2.2.2 *Un test de maîtrise*

L'importance de la validité prédictive du test de classement confère à celui-ci un statut particulier dans les typologies qu'on connaît habituellement. De fait, ils n'appartiennent à aucune des catégories dégagées par Clark (1972, 1979) Ils se distinguent des «tests pronostiques» car ils évaluent plus souvent le niveau d'habileté

que l'aptitude à apprendre et ce, bien qu'ils nous informent sur ce que pourra accomplir l'élève. Ce ne sont pas nécessairement des «tests de rendement» ou, selon la terminologie de Spolsky (1968) des «tests reliés à l'enseignement». En effet, ils ne se réfèrent pas explicitement à un apprentissage qui a déjà eu lieu. Par contre, on ne peut pas toujours les considérer comme des tests de maîtrise étant donné qu'ils cherchent souvent à identifier des forces et des faiblesses par rapport à une séquence de niveaux plutôt qu'à déterminer si l'étudiant a franchi ou non un seuil de passage préétabli. Symptôme de cette ambivalence du test de classement, on trouve au mot vedette *placement test*, dans le glossaire qui accompagne l'ouvrage de Finocchiaro et Sako (1983:305), la définition suivante: *Achievement or proficiency tests used to place students in a program or in a certain year or level of a program in a particular school.*

De fait, l'attribut «test de classement» se réfère davantage à l'usage d'un test plutôt qu'à la nature de ce qu'on cherche à évaluer. Le regroupement des étudiants peut s'effectuer autant à l'aide de tests de rendement, que de tests de maîtrise que, comme nous l'avons signalé, de tests pronostiques. Il semble qu'on puisse concevoir différents types de tests de classement selon qu'on les considère comme des tests de rendement ou des tests de maîtrise. Dans l'optique des tests de rendement, le test de classement cherche à déterminer si l'étudiant maîtrise suffisamment les éléments enseignés à un certain niveau pour pouvoir être placé au niveau suivant; le test a alors une fonction diagnostique qui peut même permettre d'établir certains objectifs devant s'intégrer au futur programme de l'élève. Par contre, le test de classement peut s'apparenter davantage à un test de maîtrise lorsqu'il prétend non pas classer l'élève en fonction d'un contenu de cours, mais en fonction d'une maîtrise générale qui peut être absente dans le cas des plus débutants ou quasi-parfaite dans le cas des plus avancés. Cette dernière approche est également celle qu'adopte Harrison (1983b:27):

> The language content of placement tests cannot be specified in detail because it must be suitable for a wide range

> of students with different learning backgrounds. The range of the students' experience is one of level as well as content, and since the intention is to separate them out into class groups, it is useful to set the tests, where possible, on an 'incline of difficulty'.

Il s'agit donc en réalité, d'un test de maîtrise qui, au lieu de ne présenter qu'un seul niveau de passage en contient autant que le nombre de niveaux désirés: débutants, faux-débutants... très avancés.

Dans le cadre de la présente recherche, nous avons préféré élaborer un test de maîtrise et ce, pour deux raisons principales. Premièrement, il en résulte un instrument qui doit être validé par rapport à une théorie plutôt que par rapport à un contenu de cours spécifique. Petersen et Cartier (1975) ne manquent pas de faire remarquer que la validité du construit et la validité concurrente, bien que plus difficiles à obtenir que la validité de contenu, permettent d'en arriver à de meilleurs tests. De cette façon, le test peut servir à un plus grand nombre d'institutions qui partagent une approche et un type de population étudiante, sans nécessairement offrir des programmes de cours tout à fait identiques. Deuxièmement, le test s'accorde mieux avec les pratiques pédagogiques qui semblent donner maintenant moins de place à la compétence grammaticale.

1.2.2.3 *La marge d'erreur*

Le degré de précision est directement relié à la variance de l'erreur acceptable laquelle dépend des besoins et des contraintes que connaissent les usagers d'un test. Ainsi, un établissement qui, à cause du nombre restreint d'élèves ou du peu de ressources disponibles, ne peut offrir plus de trois niveaux (débutant, intermédiaire et avancé), n'aura sans doute pas à recourir à un test de classement très précis - à la condition qu'on puisse traiter les cas frontières adéquatement. Il est donc souhaitable qu'on puisse disposer d'un instrument qu'on pourrait éventuellement ajuster selon des besoins ou des contraintes spécifiques et qu'on puisse suggérer quelques solutions pour le traitement des cas frontières.

Lorsqu'on juge de la pertinence d'un test de classement, il faut tenir compte de la décision à laquelle doit mener le test. À cet égard, il est nécessaire de distinguer deux types de test de maîtrise: le test de certification et le test de niveau. Le test de certification vise à déterminer si un étudiant a atteint le niveau requis pour l'admission dans un programme, l'exécution de certaines tâches professionnelles, le passage d'un cours à un autre... Il y a donc un niveau de passage unique, un seuil autour duquel se concentre le processus d'évaluation. Plus le niveau de difficulté du test s'éloigne de ce seuil, moins le test est pertinent. Par ailleurs, le test de certification mène le plus souvent à des décisions importantes et susceptibles d'avoir un impact sérieux sur l'avenir professionnel ou académique d'un candidat. Il est donc important que la marge d'erreur du test, du moins autour du niveau de passage, soit très étroite.

Contrairement au test de certification, le test de classement est un test de niveau car il n'a pas comme but de déterminer si l'étudiant a réussi mais il sert plutôt à former des groupes homogènes. D'une certaine façon, il s'agit d'un test de certification dont le nombre de seuils s'établit en fonction du nombre de groupes qu'on souhaite distinguer. Comme pour le test de certification, les variations qui n'amènent pas de passage d'un niveau à un autre ne sont pas pertinentes. Parce qu'il est généralement superflu de préciser dans quelle mesure la performance s'éloigne de ce qu'on pourrait considérer comme un niveau de passage, on peut tolérer des intervalles de confiance relativement larges. Par ailleurs, il faut noter qu'un mauvais classement attribuable à la marge d'erreur du test est rarement dramatique et irrémédiable. De fait, il est généralement possible de rectifier les erreurs du test en changeant des étudiants de groupe, en donnant des leçons de rattrapage, en développant une attitude positive chez l'étudiant... En d'autres termes, étant donné le type de décision à prendre, on peut accepter une marge d'erreur relativement grande.

1.2.2.4 *L'aspect pratique*

Le fait qu'on puisse tolérer des intervalles de confiance relativement larges affecte la fiabilité du test. On sait également que

le degré de fiabilité recherché a des implications pratiques importantes. Or, il faut garder à l'esprit le contexte habituel du test de classement: une masse d'élèves à trier rapidement. Dans ces conditions, un test qui est trop long ou trop difficile à administrer et à corriger risque de rester sur les tablettes, aussi précis soit-il. Voilà pourquoi, l'utilisation de l'ordinateur présente tant d'intérêt: on peut espérer réconcilier fiabilité et commodité.

Par ailleurs, les contraintes d'ordre pratique (commodité et économie) exercent aussi une influence sur la validité du test. Par exemple, il est indéniable que l'expression orale fournit de précieux renseignements sur le niveau général d'un étudiant. Dans le cadre du mouvement axé sur la maîtrise, l'expression orale devient même la base de l'évaluation de la maîtrise (Clifford 1980,1981). Malgré sa lourdeur administrative, l'entrevue du FSI (Jones 1978) a vite gagné de la popularité comme mesure de l'expression orale. Il n'en reste pas moins que, dans les milieux d'enseignement, on a souvent des hésitations à utiliser l'entrevue à des fins de classement. Clark (1975) reconnaît que l'entrevue directe est la méthode la plus efficace pour évaluer l'expression orale, mais il souligne les problèmes pratiques qu'elle pose et suggère qu'on mette au point des méthodes plus indirectes. Cartier (1980) décrit un test qui selon lui mesurerait des variables concomitantes à la maîtrise, au plan de l'expression orale. On a également proposé que l'entrevue orale soit administrée par des examinateurs qui n'auraient pas reçu de formation particulière (Lowe et Clifford 1980, Mattran 1977).

La somme des moyens ou de temps requise, soit pour l'administration soit pour la correction d'un test, est un facteur incontournable et extrêmement important. C'est ainsi que certaines activités à travers lesquelles on peut évaluer la maîtrise devront être mises de côté en dépit de l'apport qu'elles représentent en termes de validité de construit et de contenu. Bref, la valeur pratique d'un test de classement doit rester un souci majeur.

1.2.2.5 *Une mesure indirecte*

Le principe de l'évaluation directe est de reproduire une situation authentique c'est-à-dire une reconstruction vraisemblable

et caractéristique d'une situation que pourrait effectivement rencontrer l'apprenant. On cherche ainsi à déterminer dans quelle mesure le candidat saura faire face à des situations réelles du même type. Sans rejeter l'idée du test direct, les promoteurs de la notion de maîtrise ont remis en cause la spécificité des tâches à réaliser. De fait, l'administration d'un test de classement direct pose deux problèmes particuliers.

Tout d'abord, il faut rappeler que le test direct suppose qu'on ait établi une liste de tâches qui correspondent aux besoins communicatifs des apprenants. Or, il faut bien admettre que ces besoins ne sont pas toujours clairement identifiés. Cela est d'autant plus vrai dans un contexte scolaire où les étudiants s'inscrivent à des cours de langue avec des intérêts aussi vagues et aussi variés que l'enrichissement personnel, l'amélioration des perspectives d'emploi, l'envie de socialiser ou même la perspective d'enseigner la langue. Par ailleurs, il est courant que les besoins communicatifs auxquels on peut éventuellement rattacher des tâches spécifiques ne soient justement déterminés qu'une fois le niveau général connu c'est-à-dire une fois les épreuves de classement terminées.

L'autre problème que soulève l'évaluation directe dans la perspective du classement est relié à l'aspect prédictif que nous avons déjà touché. Le test de classement doit fournir des indications de la façon dont pourra fonctionner un apprenant dans la salle de classe selon le niveau qu'on lui aura assigné. Dans ce cas, établir des tâches spécifiques qui soient représentatives de la performance qu'on veut mesurer équivaut à dresser la liste des tâches que l'apprenant est susceptible de rencontrer dans la salle de classe. Dans les meilleurs cas, les situations de salle de classe se modèleront sur des situations authentiques. Toutefois, quoi qu'on fasse dans une salle de classe, il ne faut pas négliger le caractère foncièrement artificiel des activités pédagogiques. Par conséquent, on peut tendre vers une évaluation plus directe mais la nature même du test de classement, tourné vers la salle de classe, implique une évaluation indirecte.

1.2.2.6 *Une évaluation normée*

Pour reprendre la terminologie de Cziko (1981), le test «édumétrique» c'est-à-dire à interprétation critériée, fournit plus d'information qu'une évaluation «psychométrique» c'est-à-dire à interpréprétation normative. Cartier (1968) montre que l'évaluation critériée, plutôt que de comparer les apprenants les uns par rapport aux autres, les situe relativement à des objectifs clairement définis, à la réalisation de tâches linguistiques précises. En ce sens, une évaluation critériée est beaucoup plus difficile à réaliser (Brown 1989). On sait que le test de classement est généralement un test de maîtrise plus qu'un test de rendement; de plus, le test de classement tend à être une mesure plutôt indirecte. Dans ces conditions, on peut imaginer qu'il soit souvent difficile de concevoir un test de classement critérié.

Davies (1975) fait remarquer que le plus souvent un test sert à assigner un rang aux étudiants. Le test de classement ne fait pas exception à la règle. Ce qu'on attend de lui, c'est qu'il situe les étudiants par rapport à une dimension donnée. Les écarts entre les résultats sont rarement importants car il s'agit de remplir un certain nombre de groupes-classes. Ainsi, dans la plupart des cas, une mesure ordinale s'avère suffisante. Dans ce contexte, pourquoi faudrait-il que le test de classement soit critérié? Hormis les situations où la population étudiante s'écarte de ce qu'on trouve normalement ou les cas frontières litigieux, un test normatif peut être tout à fait satisfaisant. Après tout, le but de l'opération de classement n'est-il pas de trier les élèves?

Bien sûr, il serait souhaitable d'en arriver à une normalisation afin de permettre des comparaisons entre les programmes et entre les étudiants (à l'intérieur d'un même établissement ou entre les établissements). Cette recherche de la «commune mesure», selon l'expression de Clark (1980), pourrait éventuellement mettre fin à la balkanisation des pratiques évaluatives souvent observée dans les institutions post-secondaires (Whitny 1980, Young 1980). Il s'agit d'un souhait tout à fait légitime qui dépasse toutefois les buts du test de classement. Il faut ajouter que les techniques de standardisation que nous utilisons pour le présent test pourraient bien être adaptées en vue de la mise au point d'un instrument critérié.

1.2.2.7 *L'unidimensionalité*

Même dans le simple cas d'une mesure ordinale, il faut considérer avec beaucoup de circonspection un score qui se présente comme le total de résultats obtenus dans des épreuves de nature différente. C'est d'ailleurs une des raisons pour laquelle le débat sur le trait unitaire a pris une importance aussi grande dans le domaine du testing.

Dans les cas où la composition des scores ne permet pas de situer les sujets sur une dimension commune, il faut dresser un profil de l'étudiant. On dira que l'étudiant **A** qui a obtenu des scores élevés en discrimination auditive mais faibles en littérature et en orthographe, a un profil différent de l'étudiant **B**, qui a failli en discrimination auditive, mais bien réussi dans les deux autres épreuves. Rares sont les établissements qui pourraient tirer profit d'un tel test qui mesurerait plusieurs habiletés de façon à refléter les forces et les faiblesses de chaque élève. En effet, pour constituer des groupes-classes homogènes, il est plus convenable de trier les étudiants en fonction des résultats à un nombre restreint de sous-tests évaluant des habiletés particulières, représentatives de la maîtrise et interreliées. Il est possible que le test ne rende pas compte de la totalité de la compétence langagière. De fait, même si la performance communicative n'est pas vraiment unidimensionnelle, la fonction même du test de classement nous force à prendre une décision «unidimensionnelle». En effet, les tests de classement visent à déterminer à quel niveau d'une échelle unique appartient l'élève.

La plupart des établissements n'offrent pas de cours spécialisés portant sur les aspects spécifiques de la communication en langue seconde. Beaucoup de ceux qui le font ne peuvent se permettre d'administrer toute une batterie de tests. Ainsi, comme la décision est le plus souvent unidimensionelle, il devient important de pouvoir évaluer les étudiants en fonction de leur maîtrise générale dans la langue seconde.

1.2.3 *La nature du test*

Sans être totalement intégratif, le test ne repose pas sur un découpage systématique en unités discrètes. De fait, l'intention est de mesurer la maîtrise telle qu'elle se définit dans le courant qui a porté cette notion au premier plan. Le test n'est pas un test de compétence dans la mesure où la seule connaissance du code ne nous intéresse guère; il n'est pas non plus un test de performance car il doit permettre certaines généralisations sur ce que pourrait être la performance en de multiples situations. Nous disons donc qu'il s'agit d'un test de maîtrise. Cette maîtrise implique la connaissance du code mais aussi la mise en oeuvre de stratégies compensatoires quand la connaissance devient déficiente. Cette maîtrise se révèle à travers une performance particulière mais permet aussi le passage d'une situation de communication à une autre.

Les contraintes pratiques, c'est-à-dire la nécessité d'avoir un mode d'administration et un mode de correction commodes et économiques ont joué un rôle déterminant dans la planification du test. De même, on a dû tenir compte de l'objectif principal du projet, à savoir une comparaison entre une version «papier-crayon» et une version informatisée. Les principaux compromis que nous avons dû faire sont les suivants:

- Utilisation exclusive de questions à choix multiples: Outre ses qualités psychométriques incontestables, ce type de question est remarquablement efficace en terme de temps d'administration et de correction; de plus, c'est pour l'instant, le format le plus approprié pour les applications utilisant l'ordinateur.
- Absence d'activité de production: Malgré une validité inégalable, la correction des épreuves de production est longue, coûteuse et souvent peu fiable; par ailleurs, la production implique un aspect imprévisible dont la machine s'accommode fort mal.
- Un test indirect: La maîtrise étant une entité abstraite, nous en recherchons plutôt des indices; dans le cas

présent, comme l'étudiant est face à une machine ou face à un questionnaire, la mesure de ces indices s'effectue dans une situation tout à fait artificielle.

- Un nombre de sous-tests limité: Afin de ne pas prolonger indûment l'administration du test tout en utilisant un nombre suffisant d'items comparables, le test ne comprend que trois parties.
- L'exclusion des habiletés orales: Au plan de l'expression, l'entrevue demeure le moyen le plus efficace mais aussi le moins pratique; au plan de la compréhension, l'introduction d'une composante orale impliquait des complications techniques que nous cherchions à éviter dans le cadre de la présente recherche[4].
- Priorité aux habiletés réceptives: L'aspect imprévisible de la production posait un problème de taille dans cette recherche. Nous avons décidé de nous concentrer sur les habiletés réceptives. Il faut toutefois rappeler que de nombreuses recherches récentes ont confirmé l'interdépendance des habiletés réceptives et productives (Krashen 1981, 1983, Nagle et Sanders 1986, Faerch et Kasper 1986). Selon ces théories, la compréhension figure comme une condition préalable à la production, de sorte que la mesure de la compréhension devrait refléter les capacités d'expression.

Il est clair que ces restrictions sont incompatibles avec les caractéristiques des tests dit «communicatifs» tels que décrits par Wesche (1981) ou Swain (1984a, 1984b). Néanmoins, le test que nous avons élaboré prétend mesurer suffisamment la maîtrise en français langue seconde pour servir d'instrument de classement. Ce test est constitué de trois sous-tests. Il nous semble que les sous-tests mesurent les deux aspects essentiels de la maîtrise, du moins lorsqu'il s'agit d'assigner un groupe-classe à un apprenant:

---

[4] L'évolution rapide de nouvelles technologies permettant d'intégrer des signaux audio-visuels (sur vidéo-disque ou sous la forme de son numérisé, par exemple) offre des possibilités intéressantes quant à l'évaluation de la compréhension auditive.

- la capacité de faire des prédictions à partir d'une situation (sous-test #2) ou d'un contexte (sous-test #3);
- la capacité de reformuler une information de nature linguistique (sous-test #1) ou pragmatique (sous-test #2).

1.2.3.1 *Test de lecture*

Étant donné que les items devront être incorporés dans un test «papier-crayon» ou administrés par ordinateur, il est difficile d'évaluer la compréhension auditive. Toutefois, il est possible de juger de la compréhension de l'écrit en présentant un texte dont on mesure le degré de compréhension au moyen de questions portant sur le contenu. On aura des questions sur les relations entre les éléments tant au niveau de la phrase qu'au niveau du discours, sur des nuances stylistiques, sur le vocabulaire, sur les valeurs sociolinguistiques ou culturelles... Il est évident que cette évaluation de la compréhension par la lecture ne tient pas compte de facteurs spécifiques à la compréhension auditive tels que les stratégies d'écoute, la capacité de discrimination, les différences de registre, etc., mais elles permettent d'évaluer de façon globale le niveau de compréhension. Jafarpur (1987) a effectué une analyse de la structure factorielle de tests construits avec cette technique de «contexte court»; il a découvert que le facteur principal correspondait à une performance générale dans la langue cible.

Voici un exemple d'item où l'on demande à l'étudiant de choisir la réponse correcte:

> À partir de la semaine prochaine, les citoyens ne pourront plus garer leur voiture dans la rue, pendant la nuit. Ce règlement a pour but de faciliter le travail de déneigement au cours de l'hiver.
>
> A- Il neige depuis une semaine.
> B- Actuellement, les citoyens peuvent garer leur voiture dans la rue, la nuit.
> C- Le travail de déneigement commencera dès la semaine prochaine.
> D- On peut garer sa voiture dans la rue, pendant le jour.

La consigne est donnée en anglais afin que l'étudiant sache précisément ce qu'on attend de lui. On pourrait aisément traduire les directives de ce sous-test pour viser des étudiants dont la langue maternelle ne serait pas l'anglais puisque les choix de réponse sont dans la langue cible. Cette dernière caractéristique implique aussi que certains problèmes de compréhension pourront survenir dans la lecture des choix de réponse. Ceci offre plus de liberté dans la rédaction des items et permet de mieux contrôler le niveau de difficulté. Enfin, il faut noter qu'on pourrait avoir recours à une variante de ce genre d'items où l'on utiliserait un texte plus long sur lequel on poserait plusieurs questions dont le niveau de difficulté varierait.

1.2.3.2 *Choix de l'énoncé qui convient*

Il est difficile dans le type de test que nous proposons, d'évaluer les productions spontanées ou les énoncés produits dans un contexte donné. Malgré des développements intéressants dans le domaine de l'intelligence artificielle et de l'analyse automatique du discours (pour une synthèse de la question, voir Bonnet 1984), on est encore loin du jour où la machine pourra jouer le rôle d'un locuteur sensible aux particularités d'une situation de communication. Newsham (1989) signale que même dans le cadre d'une entrevue il est difficile de faire varier les paramètres de la situation de communication pour vérifier comment le sujet tient compte de ces paramètres dans ses productions. Howard (1980) et Raffaldini (1988) indiquent qu'on peut évaluer la capacité de l'élève à produire des énoncés corrects et appropriés en lui soumettant une série d'énoncés parmi lesquels il doit choisir celui qui correspond le mieux à la situation décrite.

À titre d'exemple, on peut citer l'item suivant sur lequel se modèlent ceux de la deuxième partie:

> *You are in the train. You do not know the passenger who is sitting beside you and you wonder if you may*

smoke. The person is a man, about 50 years old; he is reading a magazine. To inquire, which question would you use?

A- Tu veux que je fume?
B- Auriez-vous l'extrême obligeance de me permettre de fumer?
C- Est-ce que cela vous dérange si je fume?
D- Il faut que je fume.

Dans cet exemple, on décrit la situation en utilisant la langue maternelle du sujet. Hormis l'usage d'images, qui risqueraient d'être ambigües, on imagine mal comment il pourrait en être autrement. Les items de cette section renseignent sur la capacité de l'étudiant de repérer un énoncé qui soit à la fois correct et approprié dans une situation donnée. Avec ce sous-test, on fait donc une place à la dimension socio-culturelle (Condon 1975) de même qu'à la variation sociale (Duran 1984). On reconnait deux des trois critères retenus par Morrow (1982) pour l'évaluation de l'expression orale. En effet, Morrow distingue entre *appropriacy, accuracy et fluency*. Notons que le dernier critère ne peut servir, compte tenu du genre de test que nous développons. Malgré certaines limites sérieuses (Cazabon-Size et Cazabon 1986), ce genre d'item qu'on ne retrouve généralement pas dans les tests standardisés, permet donc d'évaluer à la fois la compétence grammaticale et la compétence sociolinguistique.

1.2.3.3 *Test de phrases lacunaires*

Depuis que Oller (1979) a établi la typologie des tests de closure et en a fait les louanges en prétendant qu'ils constituaient une mesure intégrative propre à activer la «grammaire de l'expectative», une littérature abondante s'est développée autour de cette question. Cole (1981) fait la synthèse de plusieurs études qui ont confirmé la validité concurrente des tests de closure; les corrélations sont particulièrement remarquables avec les épreuves de grammmaire et de vocabulaire. Ces résultats ont été par la suite corroborés par des recherches complémentaires (Hinofotis 1980, Stansfield 1982, Hanania et Shikhani 1986). Certains (Lee

1985) parlent même de validité du construit alors que d'autres (Farhady 1983c) s'interrogent sur ce que mesurent réellement les tests de closure. De fait, le test de closure a connu un tel succès qu'aujourd'hui il fait souvent office de test de classement.

Plusieurs modifications au test de closure traditionnel ont été proposées. Certaines suggestions comme la technique de *Clozentropy* (Brown 1980), la technique de *ClozeElide* (Manning 1985) ou le test-C (Klein-Bradley et Raatz 1984) se distinguent par leur originalité. On a observé que l'efficacité du test de closure pouvait être améliorée en sélectionnant les effacements plutôt qu'en procédant de façon aléatoire (Oller et Inal 1975, Bondaruk *et al.* 1975, Bachman 1982). On a aussi noté que des formules où l'on acceptait tout mot acceptable dans le contexte ou des versions adoptant le format des questions à choix multiple permettaient d'éviter les écueils de la correction par le mot exact (Brown 1980). Des formules à choix multiples se sont avérées tout aussi valables et beaucoup plus pratiques quand il s'agit de mesurer la maîtrise générale (Jockens et Montens 1988, Jonz 1976).

Toutefois, plusieurs recherches ont mis en doute la valeur intégrative du test de closure et démontré que cette technique ne permettait que d'évaluer des facteurs de bas ordre (Alderson 1980, 1981, Connors et Toker 1984, Porter 1983). Selon ces études, le contexte qu'utilise effectivement le sujet pour compléter le texte, ne dépasse pas quelques mots de sorte que les indices discursifs jouent un rôle secondaire. Dans cette perspective, on peut donc penser qu'un simple exercice de phrase lacunaire fait appel à plusieurs stratégies et de connaissances communes. L'exercice de phrases lacunaire ne peut se substituer totalement au test de closure, mais en ce qui a trait à la maîtrise générale, il semble apporter une information du même ordre.

La plupart des tests standardisés ont recours à ce type d'items. Beaucoup de manuels de testing, publiés durant la période «psychométrique-structurale», accordent une place prépondérante aux phrases lacunaires (Lado 1961, Harris 1969, Valette 1977). Par rapport au test de closure, ce format a l'avantage de préserver le

principe de l'indépendance des items. L'analyse statistique du test s'en trouve donc grandement facilitée. Par conséquent, les items du troisième sous-test ont la forme suivante:

> En plus de vous faire voyager plus rapidement, nous vous enverrons où les autres compagnies ne font _____ pas escale.
>
> A- même
> B- surtout
> C- rarement
> D- heureusement

1.2.3.4 *Les niveaux de maîtrise*

Nous tenterons de tenir compte de la majorité des utilisateurs d'un test de classement en établissant sept niveaux différents:

niveau 1 → Débutants
niveau 2 → Faux débutants
niveau 3 → Intermédiaires faibles
niveau 4 → Intermédiaires moyens
niveau 5 → Intermédiaires forts
niveau 6 → Avancés
niveau 7 → Très avancés

On notera que cette division suit de près l'échelle de niveaux établie par l'*American Council on the Teaching of Foreign Languages* (*ACTFL*, Byrnes et Canale 1987) ou même par le *English Language Testing Service* (*ELTS*, Seaton 1983). De fait, la description des niveaux suit l'esprit d'un projet de normalisation internationale connu sous le nom de *The Common Yardstick* (Educational Testing Servive 1978, Clark et Clifford 1988). Ces échelles comprennent neuf niveaux. En ce qui nous concerne, on peut fusionner les deux premiers niveaux étant donné le bagage que possèdent habituellement les apprenants de la population visée; on peut également fusionner les deux derniers niveaux puisque, théoriquement, les étudiants du niveau le plus élevé n'ont plus besoin de cours de langue seconde. De plus, afin de nuancer l'estimation du niveau, on pourra ajouter des catégories mitoyennes: par exemple, on pourra classer un apprenant au niveau «Faux débutant +».

②

LES PRINCIPES DU TESTING ADAPTATIF

Avec le développement technologique dans le domaine informatique et la mise au point de techniques docimologiques toujours plus raffinés, on voit apparaître de plus en plus de tests qui utilisent l'ordinateur. De fait, l'ordinateur présente deux caractéristiques qui rendent son utilisation particulièrement intéressante pour le testing. D'une part, il offre la possibilité de branchements multiples: le logiciel peut donc «prendre des décisions» au cours de l'administration du test. D'autre part, sa capacité de traitement numérique permet d'exécuter très rapidement des calculs complexes dont on peut utiliser les résultats sur le champ. Le testing adaptatif[1] apparaît comme le moyen idéal d'exploiter à fond ces deux caractéristiques intéressantes de la machine.

2.1 Caractéristiques des tests adaptatifs

2.1.1 *Le concept de testing adaptatif*

Larson et Madsen (1985) font remarquer que l'utilisation de l'ordinateur pour l'évaluation en langue seconde peut s'inspirer des didacticiels d'enseignement mais qu'elle est plutôt appelée à s'en distinguer et à appliquer de plus en plus la notion de testing adaptatif. Cette notion se comprend aisément dans le cadre

[1] À défaut de terme français attesté, nous nous rallions à une tendance selon laquelle le néologisme «testing adaptatif» devrait correspondre à l'équivalent anglais *adaptive testing*.

d'un test de classement. En effet, de par sa nature, le test de classement implique qu'on l'administre à un groupe dont le niveau d'habileté varie considérablement. On doit retrouver dans le test de classement des questions qui s'adressent à chaque niveau. On accepte donc que l'étudiant débutant soit confronté à des questions généralement beaucoup trop difficiles et qu'inversement, l'étudiant très avancé trouve le test extrêmement facile. De fait, peu importe le niveau de l'étudiant, la plupart des questions d'un test de classement sont soit trop faciles soit trop difficiles. Outre les effets psychologiques qu'on peut imaginer (frustration, abandon, inattention...), cette situation affecte la qualité de la mesure. En effet, lorsque la probabilité de réussite ou d'échec à un item devient trop grande, cet item apporte peu d'information. On comprend alors l'intérêt d'un test au cours duquel l'apprenant serait soumis à des items adaptés à son niveau, c'est-à-dire ni trop difficiles, ni trop faciles. Pour expliquer cette notion de testing adaptatif, Wainer (1983) a recours à une analogie avec la course à obstacles: au cours de l'épreuve, on tente de placer des barrières que le coureur a autant de chance de franchir que de faire tomber. Weiss et Kingsbury (1984:361) définissent ainsi le testing adaptatif: *Adaptive testing is a process of test administration in which test items are selected for administration on the basis of the examinee's response to previously administered items.* Comme le souligne Anastasi (1982:304), ce n'est pas le nombre de questions qui importe, mais le niveau où se déroule le test: *The individual score is based, not on the number of items answered correctly but on the difficulty level and other psychometric characteristics of those items.* Il en résulte que dans ce test «sur mesure», les items administrés varient nécessairement d'un apprenant à l'autre.

Dans le domaine de la didactique des langues, les récentes techniques d'entrevue pour l'évaluation de l'expression orale nous ont habitués à ce type d'épreuve que l'examinateur mène en fonction des hypothèses qu'il construit quant au niveau réel du sujet (Wilds 1975). Dans le domaine de la psychologie, déjà au début du siècle, Binet (1909) reconnaissait le principe du testing adaptatif dans la mesure de l'intel-

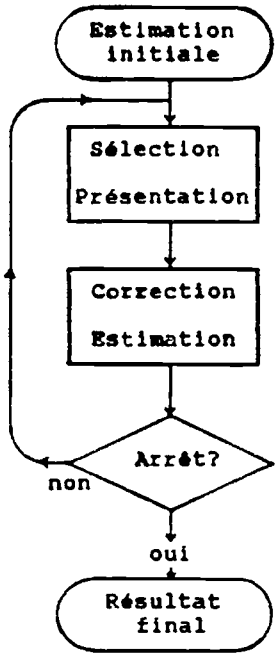
ligence. Le test se modelait selon les réponses fournies et comprenait les étapes que distinguent Kreitzberg et al. (1978) dans le déroulement d'un test adaptatif. Ceux-ci y reconnaissent quatre étapes:

- Obtenir une estimation initiale du niveau d'habileté: Il peut s'agir de la moyenne de la population, du résultat d'une épreuve antérieure, d'une approximation quelconque... Cette première estimation sert à amorcer la procédure.
- Déterminer un item approprié: On choisit parmi les items qui n'ont pas encore été présentés celui qui tient le mieux compte de l'estimation du niveau du sujet. On affiche alors l'item à l'écran et on demande au sujet de répondre.
- Corriger et réviser l'information: Dès que le sujet a répondu, on enregistre la réponse et cette information est utilisée pour recalculer le niveau d'habileté qui servira de base au choix du prochain item.
- Arrêter lorsque l'estimation est assez précise: Si on juge que l'estimation est assez fiable, on interrompt l'administration. Dans le cas contraire, on retourne à la deuxième étape.

Le déroulement peut se représenter selon l'organigramme de la figure 2.1.

Plus que les effets psychologiques chez l'étudiant qui fait le test, c'est l'efficacité de la procédure qui a surtout retenu l'intérêt de ceux qui s'y sont intéressés. Green (1983a), souligne l'importance de la notion d'information dans les procédures de testing adaptatives les plus récentes. Il s'agit essentiellement de maximiser cette information c'est-à-dire d'obtenir des données qui permettront d'arriver à une estimation qui soit le plus près possible du niveau réel de l'étudiant et ce en utilisant un nombre minimum d'items.

FIGURE 2.1
Schéma du déroulement d'un test adaptatif



Toutefois, outre l'avantage du point de vue psychométrique, Green (1983b) relève quelques autres avantages que peut présenter la présentation par ordinateur:

- Comme aucun document ne circule (questionnaire, feuille de réponse ou grille de correction) et que les étudiants ont des tests différents, on préserve la confidentialité du test.
- Puisque l'étudiant répond au clavier, on évite les feuilles de réponses parfois illisibles et la correction des réponses est immédiate.
- Le test s'administrant de façon individuelle, l'étudiant peut le faire à son propre rythme, sans devoir subir les contraintes de temps de l'administration en groupe.

- On évite la frustration et on stimule l'intérêt de l'étudiant qui répond à des questions correspondant à son niveau d'habileté.
- La présentation par ordinateur permet de créer de nouveaux types d'items en exploitant les possibilités graphiques, le clavier, le son...

Dans la dernière partie de cette recherche, nous considérerons plus en détail les avantages et les inconvénients que présente un test adaptatif en langue seconde par rapport aux tests traditionnels. Toutefois, avant d'y arriver, il nous a fallu, élaborer un tel test. Nous avons donc créé ce qui constitue les deux composantes essentielles de tout test adaptatif: une procédure de sélection des items et une banque d'items.

2.1.2 *La procédure de sélection des items*

L'étudiant à qui on administre un test conventionnel devra, à moins que sa fantaisie ne lui suggère un ordre différent, répondre d'abord au premier item qui apparaît dans le questionnaire (ou sur la bande) puis passer au second et ainsi de suite. Par contre, si le test est administré avec un ordinateur, il n'est pas nécessaire que l'ordre linéaire de présentation des items soit respecté. Le concepteur d'un test informatisé peut programmer diverses stratégies de sélection des items.

2.1.2.1 *Administration linéaire*

Ce type de test informatisé ne peut pas être considéré comme un test adaptatif puisqu'il s'agit le plus souvent de la simple transposition à l'écran d'un test conventionnel. On présente les items selon un ordre pré-déterminé en tenant compte des limites inhérentes à ce mode de présentation et des possibilités qu'il offre. Dans la typologie de Bunderson, Inouye et Olsen (1989), ce type de test correspond à la première génération des tests informatisés.

2.1.2.2 *Sélection aléatoire*

Ce type de test décrit par Lord (1977b) s'appuie sur la théorie des tests aléatoirement parallèles (Lord et Novick 1968:chap 11). Le programme choisit au hasard parmi un ensemble d'items homogènes, un nombre pré-déterminé d'items. Cette technique a été utilisée par Emerson (1974) afin de générer des tests différents à chaque administration. Toutefois, bien qu'aucun sujet ne reçoive le même test, cette stratégie de sélection ne constitue pas véritablement un test adaptatif.

2.1.2.3 *Branchement en fonction du contenu*

La sélection s'effectue en fonction de la nature des items présentés et non pas de la réponse du sujet. On peut ainsi limiter le nombre d'items se rapportant à un aspect spécifique ou programmer l'exclusion d'une classe d'items à la suite d'un item particulier.

2.1.2.4 *Test à plusieurs étapes*

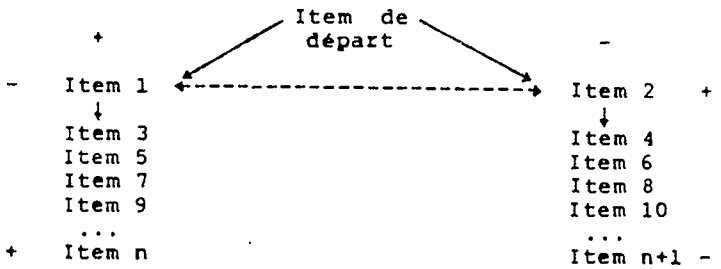
Dans sa forme la plus simple, le test à étapes (Betz et Weiss 1974), comprend un tronc commun qu'on administre à tous les sujets. Selon le niveau d'habileté calculé, le programme choisit alors une séquence particulière d'items. On peut reprocher à ce type de test que la fiabilité du résultat final dépende beaucoup de la décision prise après la première étape. Cleary *et al.* (1968) observent que la qualité de la mesure obtenue après l'évaluation préliminaire varie selon la façon dont les items ont été sélectionnés. Selon eux, on peut réduire le nombre d'erreurs de classement si on détermine la séquence en révisant la catégorie où se trouve l'étudiant après chaque item.

2.1.2.5 *Test flexilevel*

Conçu par Lord (1971), d'abord pour les tests «papier-crayon», ce type de sélection se prête facilement à une adminis-

tration informatisée. Les items sont rangés dans deux séries: la série de gauche comprend des items ordonnés du plus facile au plus difficile et dont le niveau de difficulté est supérieur à la moyenne; dans l'ordre inverse, on retrouve dans la série de droite, des items dont le niveau de difficulté est inférieur à la moyenne (figure 2.2). On administre d'abord un item de départ de difficulté moyenne; si la réponse est exacte, on choisit le prochain item dans la série de gauche, sinon on choisit dans la série de droite. Comme le souligne Séguin (1976), le test *flexilevel* s'avère surtout efficace lorsque la gamme des niveaux à l'intérieur d'un groupe est très étendue.

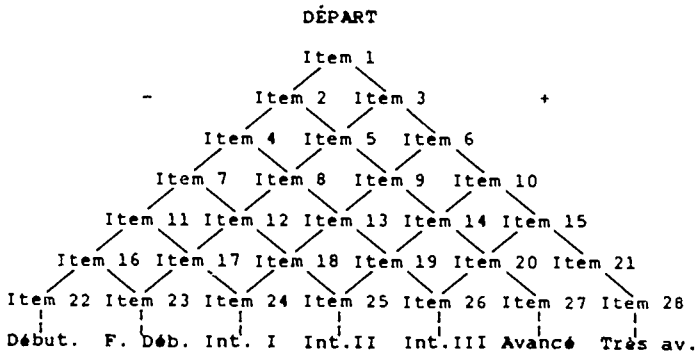
FIGURE 2.2
Déroulement d'un test *flexilevel*



2.1.2.6 *Test pyramidal*

Larkin et Weiss (1974) désignent ainsi ce type de sélection à cause de la hiérarchie dans laquelle prennent place les items. En effet, ceux-ci sont ordonnés dans une structure arborescente. Si le sujet fournit une réponse exacte il est dirigé vers l'item plus difficile; sinon on lui soumet l'item plus facile. En distinguant les sept niveaux que nous avons décrits, la structure se présente comme la figure 3. On peut se demander si cette procédure utilise efficacement les items disponibles. Par ailleurs, afin d'améliorer la fiabilité, on peut subtituer à chaque item qui occupe un noeud de l'arbre, un groupe d'items. Par exemple, la réussite de trois items d'un groupe de cinq dirige le sujet vers le groupe d'items plus difficiles.

FIGURE 2.3
Déroulement d'un test pyramidal



2.1.2.7 *Test par stratification*

Le test *stradaptive* (Vale et Weiss 1975) consiste à diviser le continuum que représente l'habileté en un certain nombre de strates. À chacune, sont associés des items dont le niveau de difficulté correspond à celui de la strate. La structure des items d'un test à sept strates se présente sous la forme d'une matrice dont la longueur varie selon le nombre d'items disponibles (figure 4). Il est toutefois possible qu'un item particulièrement efficace se retrouve dans plusieurs strates adjacentes. Si l'étudiant répond correctement, on lui soumet le prochain item de la strate de niveau supérieur; sinon on lui présente celui de la strate de niveau inférieur.

FIGURE 2.4
Déroulement d'un test par stratification

Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7
Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14
Item 15	Item 16	Item 17	Item 18	Item 19	Item 20	Item 21
Item 22	Item 23	Item 24	Item 25	Item 26	Item 27	Item 28
Item 29	Item 30	Item 31	Item 32	Item 33	Item 34	Item 35
Item 36	Item 37	Item 38	Item 39	Item 40	Item 41	Item 42
Item 43	Item 44	Item 45	Item 46	Item 47	Item 48	Item 49
Début.	F. Déb.	Int. I	Int.II	Int.III	Avancé	Très av.

2.1.2.8 *Test par correspondance*

Lord (1970) indique qu'on peut déterminer le niveau où un item est le plus efficace, c'est-à-dire où il fournit le maximum d'information. Ce niveau est calculé sur la même échelle que le niveau de l'étudiant. La procédure consiste donc à choisir, parmi les items qui n'ont pas été présentés, celui qui est le plus efficace compte tenu de l'estimation du niveau de l'étudiant. À moins que le sujet ait l'occasion de deviner la réponse, il s'agira idéalement d'un item pour lequel il y a autant de chances d'obtenir une réponse exacte qu'une réponse inexacte.

La plupart des stratégies de sélection des items supposent un classement des items selon leur difficulté relative. Si l'échantillon qui sert à la calibration est représentatif, il est possible d'utiliser des indices de probabilité pour représenter la difficulté des items. Par ailleurs, si les quatre premiers types de tests informatisés peuvent être corrigés en comptant le nombre de réponses exactes, l'estimation de l'habileté à un test pyramidal ou par stratification consiste à attribuer un niveau plutôt qu'un score. Quant à la dernière stratégie de sélection des items, le test par correspondance, elle implique que la difficulté des items et l'habileté des sujets soient mesurées à l'aide d'une échelle commune. Cela pose ainsi un problème considérable dans le cadre de la théorie classique (Lord et Novick 1968) particulièrement lorsque le hasard intervient (avec des questions à choix multiple) ou que les items ne discriminent pas également.

Par ailleurs, toute stratégie de sélection des items doit comprendre un critère qui permet d'interrompre la procédure. Dans le cadre de la théorie classique, le nombre d'items administrés demeure le critère le plus commode. Le test est terminé dès qu'un nombre pré-déterminé d'items a été présenté (ou que tous les items ont été présentés). Toutefois, la mise au point de techniques de mesure reliées à la théorie du trait latent permet maintenant de considérer la somme d'information recueillie au cours du test. Ainsi, quand on a accumulé une certaine quantité d'information, on arrête le test. La théorie du trait latent dont nous verrons les fondements

dans la section suivante, offre également une solution au problème que pose l'adoption d'une échelle commune entre la difficulté des items et l'habileté des sujets.

2.1.3 *Les banques d'items*

L'efficacité de la stratégie de sélection des items dépend de la qualité des items qui sont disponibles. Les items doivent être reliés au contenu qu'on cherche à mesurer, ils doivent servir à départager la population et on s'attend généralement à ce qu'ils couvrent une gamme suffisamment large de niveaux d'habileté. Au cours de l'administration d'un test adaptatif, le programme aura accès aux items qui sont rangés et répertoriés dans une (ou plusieurs) banque(s).

Millman et Arten (1984) définissent ainsi la notion de «banque d'items»: *a relatively large collection of easily accessible test questions*. Puisque la banque d'items permet d'obtenir un très grand nombre de tests différents, on comprend qu'elle soit une composante essentielle dans un système de testing adaptatif. Toutefois les «banques d'items» ne sont pas nées avec le concept du testing adaptatif. Il s'agissait d'abord de réunir un ensemble d'items partageant certaines caractéristiques quant à leur forme et à leur contenu et d'accéder à ces items selon les besoins (Choppin 1968). Outre ses applications dans le domaine du testing adaptatif, la banque d'items se prête à une variété d'usages:

- pour garder à jour les renseignements relatifs à l'utilisation des items;
- pour obtenir des versions parallèles ou équivalentes d'un test;
- pour abréger ou allonger un test;
- pour regrouper des items portant sur un contenu particulier;
- pour tester à un niveau de difficulté spécifique.

Comme le fait remarquer Theunissen (1987), la mise sur pied d'une banque d'items est un élément important pour «optimiser» un test c'est-à-dire utiliser le minimum d'items pour obtenir le maximum d'information dans un domaine particulier au(x) niveau(x) d'habileté pré-établi(s).

Selon le nombre d'items, le type de standardisation, l'information à emmagasiner et le système de gestion, l'élaboration de la banque d'items pourra faire appel à une technologie relativement simple (DesBrisay 1988, Nitko et Hsu 1984) ou beaucoup plus complexe (Henning 1986, Wright et Bell 1984). Dans tous les cas, le système impose une relative uniformité quant à la forme des items de sorte que beaucoup de banques se limitent, par exemple, à des questions à choix multiple. De plus, les items doivent pouvoir être traités isolément: ainsi, une série de tests de closure est difficile à intégrer dans une banque d'items. Enfin, il est essentiel que les items soient homogènes du point de vue de leur contenu c'est-à-dire qu'ils mesurent un domaine commun. Si on ne peut assurer cette comparabilité de contenu, il est possible de constituer plusieurs banques inter-reliées ou d'identifier des sous-ensembles de la banque par des mots clés.

L'ensemble des items doit être chapeauté par un système de gestion grâce auquel l'usager pourra accéder à la banque pour effectuer les trois opérations suivantes:

- **Retrouver**: L'information contenue dans la banque doit être rapidement disponible, en tout ou en partie, tant pour la construction d'un test que pour la consultation.
- **Corriger**: On doit pouvoir corriger une erreur de frappe, modifier un distracteur, enregistrer les résultats d'une recalibration...
- **Ajouter**: La structure doit être ouverte de façon à permettre l'addition de nouveaux items à la banque.

Les développements de la micro-informatique rendent ces opérations de plus en plus faciles. Plusieurs logiciels courants conçus pour la gestion de bases de données, peuvent effectuer ces fonctions (Henning 1986).

Le nombre et le contenu des champs que contient chaque fiche d'item peuvent varier selon les types de banques mais on devra nécessairement y retrouver l'information suivante:

- Un code d'identification: il sert à identifier l'item et, s'il y a lieu, la banque auquel il appartient de même que le moment où il a été inséré.
- La question: il s'agit du texte qu'on soumet (à l'écran, sur papier, sur bande ou sur disque) et à partir duquel le sujet doit répondre.
- La réponse: on entre la/les réponse(s) correcte(s).
- Des indices statistiques: si on utilise les indices classiques, on inscrit la probabilité de réponse correcte (ou le niveau de difficulté), la corrélation bisérielle (ou point-bisérielle) ou un indice de discrimination quelconque; si on utilise la théorie du trait latent, on inscrit les résultats de la calibration (indice de difficulté et s'il y a lieu, indices de discrimination et de hasard).

On pourra aussi ajouter, selon les besoins, des renseignements complémentaires:

- Les mots clés: dans les cas où la banque contient plusieurs sous-ensembles.
- Les options de réponses: cette information doit suivre les questions à choix multiple.
- Le texte complémentaire: ce peut être le passage sur lequel porte une question ou tout autre texte pertinent.
- Les données sur l'expérimentation: on peut indiquer la/les date(s) de l'expérimentation de l'item et le nombre de sujets impliqués.

- L'adéquation: ce peut être un indice du degré d'adéquation de l'item en fonction des autres items et du modèle retenu.
- Tout autre renseignement jugé important.

Une banque peut regrouper un très grand nombre d'items et de champs d'information. Cependant si la banque doit servir au testing adaptatif, il est important de ne pas surcharger les fiches afin d'assurer des temps d'accès raisonnables et de limiter l'espace requis pour emmagasiner les données. Pour les mêmes raisons, il faut veiller à ne pas multiplier le nombre d'items. Un test administré par micro-ordinateur de faible puissance auquel on adjoindrait une banque d'au-delà de 300 items risquerait de fournir un rendement médiocre. Ces contraintes sont beaucoup moins sérieuses lorsque la banque sert comme simple outil de référence ou pour produire des questionnaires de test.

Quoique la gestion de vastes banques de données ne pose généralement pas de problème technique, leur développement est souvent difficile du fait qu'il devient impossible d'administrer tous les items à une même population, dans des conditions stables. Si on estime la difficulté des items à partir de la probabilité de réponse correcte, la marge d'erreur de ces indices peut s'avérer assez importante surtout si les échantillons de sujets ne sont pas comparables. Des techniques d'ancrage mises au point dans le cadre de la théorie du trait latent (Henning 1987:chap 9, Vale 1986) permettent maintenant, sous certaines conditions, d'exprimer sur une même échelle, la difficulté d'items expérimentés avec des échantillons différents.

2.2 La théorie du trait latent

Dans le contexte de l'évaluation de la langue seconde, on a souvent remis en question le bien-fondé de la théorie classique. Nous avons mentionné quelques insuffisances de la théorie classique notamment en ce qui a trait à l'établissement d'une échelle de difficulté des items. D'une part, on souhaite obtenir une corres-

pondance directe entre la difficulté des items et le niveau d'habileté des sujets. D'autre part, il est essentiel de pouvoir comparer entre eux, tous les items qui peuvent être intégrés à la banque. La théorie du trait latent offre une solution à ces problèmes. Il est inutile de rappeler les fondements de la théorie classique dans le domaine psychométrique. On trouve d'excellents ouvrages d'introduction (Gulliksen 1950, Allen et Yen 1979, Bernier 1985) de même que des ouvrages qui approfondissent des notions de base telles que le concept de score véritable (Lord et Novick 1968) ou de fiabilité (Cronbach 1970). Toutefois, en raison de sa nouveauté, de la controverse qu'elle suscite et de l'intérêt qu'elle présente en testing adaptatif, la théorie du trait latent mérite une attention particulière.

2.2.1 *Les différents modèles*

Comme le fait remarquer le pionnier de la théorie du trait latent, Frederick Lord (1980:7) cette nouvelle approche en psychométrie est l'approfondissement de certains concepts de la théorie classique plutôt qu'une rupture avec cette dernière. La théorie du trait latent doit sa dénomination au fait qu'elle postule qu'un test est le reflet d'une caractéristique que l'on cherche à mesurer. Cette caractéristique, que ce soit l'intelligence, le vocabulaire ou la maîtrise de la langue seconde, se nomme le «trait». On le dit «latent» du fait qu'il n'est pas observable. Comme le soulignent Hambleton et Cook (1977), c'est au moyen d'une fonction mathématique que l'on peut relier la performance lors d'un test au trait sous-jacent. Il est donc essentiel, dans le cadre de la théorie du trait latent, que le test ne mesure qu'un seul trait c'est-à-dire que le test soit unidimensionnel. Si la théorie classique suppose une certaine unidimensionalité, notamment en ce qui concerne la notion de fiabilité ou l'interprétation des scores, cette exigence devient prépondérante dans le cadre de la théorie du trait latent. L'unidimensionalité implique que le test mesure le même trait dominant à tous les niveaux d'habileté et que la fonction mathématique qui relie la performance au trait est identique pour tous les sous-ensembles de la population. La théorie du trait latent porte aussi la désignation «théorie de réponse aux items». Séguin et Auger (1986:8)

signalent l'acception que prend le terme «item» dans la perspective d'une théorie qui postule l'unidimensionalité: «définition opérationnelle d'un aspect particulier de l'habileté mesurée».

La fonction mathématique entre le trait et l'habileté se représente sous la forme d'une courbe caractéristique d'item que Hambleton et Swaminathan (1985:25) définissent ainsi: "An item characteristic curve (ICC) is a mathematical function that relates the probability of success on an item to the ability measured by the item set or test that contains it". Analysant le cas de tests à réponses ouvertes (aucun hasard) corrigées de façon dichotomique (exact ou inexact), Lord (1953) a démontré le premier que cette fonction non linéaire se définissait selon la formule 2.1:

$$P_i(\theta) = \int_{-\infty}^{a_i(\theta - b_i)} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \qquad (2.1)$$

$P_i(\Theta)$ est la probabilité qu'un sujet donné dont le niveau d'habileté est Θ, réponde correctement à l'item i. z est un écart normal d'une distribution dont la moyenne est b_i et l'écart type $1/a_i$. Le symbole a désigne un paramètre qui représente la discrimination de l'item alors que *b* est un paramètre qui en représente la difficulté; les valeurs que peuvent prendre ces paramètres sont théoriquement infinies. En pratique, les valeurs de *a* oscillent entre 0 et 2 et indiquent la pente de la courbe au point où $\Theta = b$. Plus *a* augmente, plus la pente est forte et mieux l'item discrimine. La valeur de *b* s'exprime sur la même échelle que Θ, les deux indices étant transformés selon l'échelle d'une courbe normale c'est-à-dire de façon à ce que leur moyenne soit de 0 et leur écart type de 1. La valeur de *b* varie donc habituellement entre -2, pour un item très facile, à 2, pour un item très difficile. Les valeurs de *a* et de *b* doivent être estimées pour chaque item. Parce qu'on considère à la fois la difficulté et la discrimination, cette formule définit la fonction ogivale normale à deux paramètres.

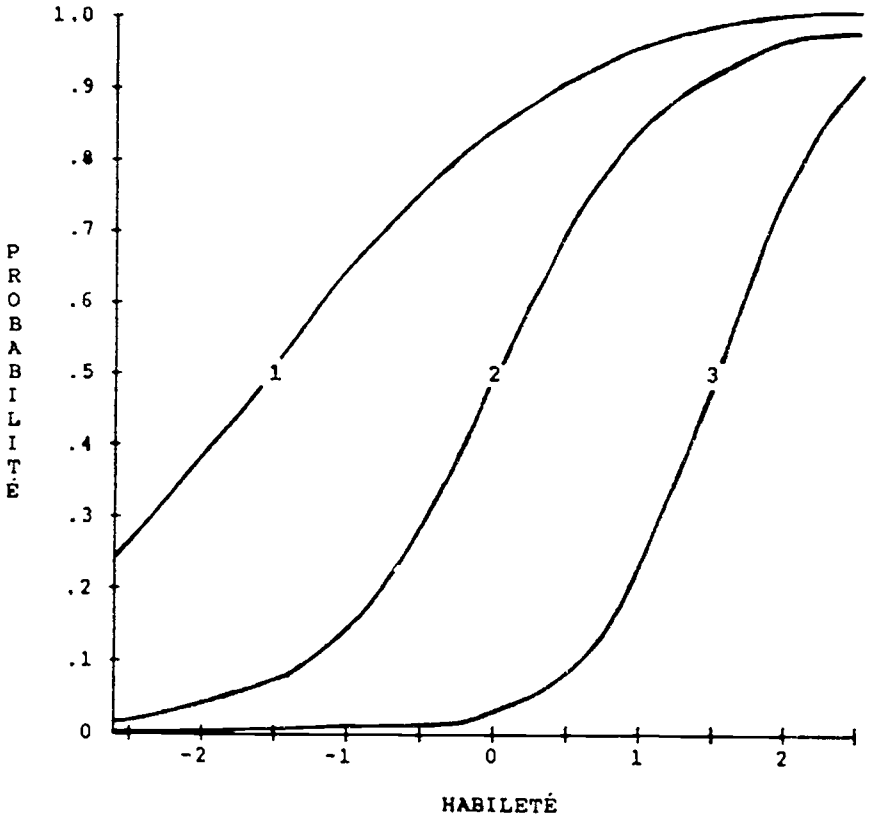
Afin de rendre la formule mathématiquement plus facile à manipuler, Birnbaum (1968) a proposé une série de formules que l'on peut substituer à la formule originale. Le modèle logistique à deux paramètres se définit donc par la formule 2.2:

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad (2.2)$$

L'utilisation d'une constante D = 1.7 dans la fonction logistique permet de reconstituer la fonction ogivale normale.

La figure 2.5 illustre quelques exemples de courbes caractéristiques d'items utilisant un modèle à deux paramètres.

FIGURE 2.5
Courbes d'un modèle à 2 paramètres

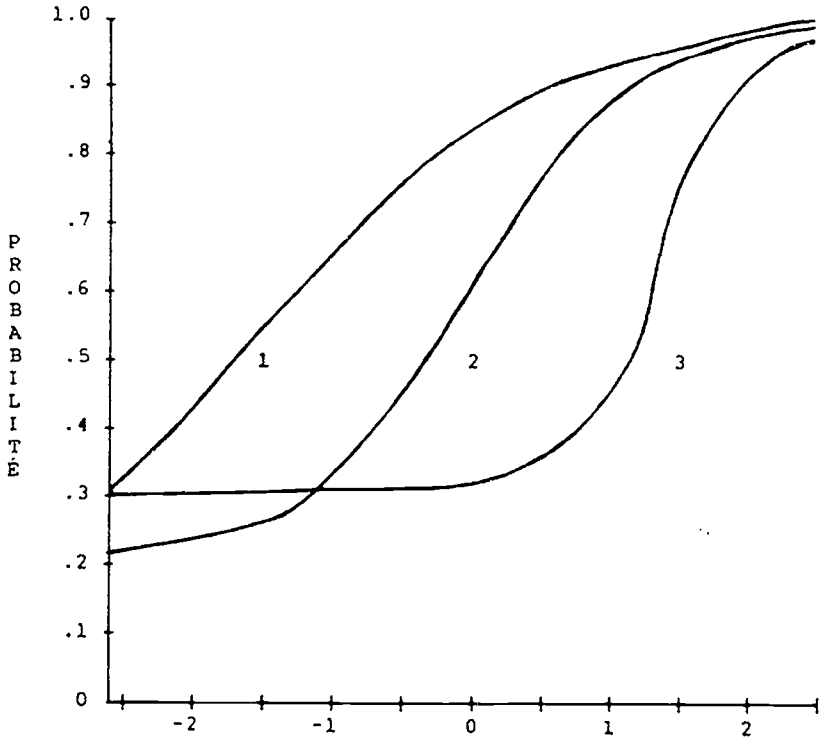


1 : a = .6 b = -1.5
2 : a = 1.0 b = 0
3 : a = 1.4 b = 1.5

Quand on utilise des questions à choix multiple, on introduit un facteur de hasard dont l'influence s'exerce plus particulièrement chez les sujets au bas de l'échelle d'habileté.

Dans ce type de test, on tient compte du hasard en ajoutant un paramètre supplémentaire *c*.

FIGURE 2.6
Courbes d'un modèle à 3 paramètres



1 : a = .6 b = -1.5 c = .1
2 : a = 1.0 b = 1.0 c = .2
3 : a = 1.4 b = 1.5 c = .3

Le paramètre *c* étant un indice de probabilité, sa valeur se situe entre 0 et 1. On peut penser que $c = 1/N$, où N correspond au

nombre d'options offertes. En pratique, les distracteurs n'exercent pas tous la même influence de sorte que c doit être estimé et pris en considération dans le calcul de la probabilité de réponse exacte. On définit alors un modèle à trois paramètres selon la formule 2.3:

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad (2.3)$$

Ainsi qu'on peut l'observer dans la figure 2.6, c définit l'asymptote de la courbe caractéristique de l'item puisque le modèle à trois paramètres prévoit que $P_i > 0$.

Le modèle à trois paramètres est lourd et requiert l'estimation de trois variables. Le statisticien danois G. Rasch (1960) a donc proposé une simplification du modèle logistique, soit un modèle à un paramètre où seule la difficulté varie. Le modèle a par la suite connu beaucoup de succès notamment sous l'impulsion des travaux de Wright et Panchapakesan (1969) et Wright et Stone (1979). Le modèle de Rasch présuppose que le hasard n'intervient pas et que tous les items discriminent également de sorte que $a = 1$ et $c = 0$. La figure 2.7 montre que les items se distinguent essentiellement par leur position sur l'échelle de l'habileté. En fixant a et c, l'estimation des paramètres (la calibration) est grandement facilitée. De plus, on peut maintenant considérer le score obtenu à un test (le nombre de réponses exactes) comme l'estimation la plus juste de l'habileté du sujet. La procédure d'estimation des paramètres des items (la calibration) reste complexe, nécessite un grand nombre de sujets et doit s'effectuer à l'aide d'un ordinateur.

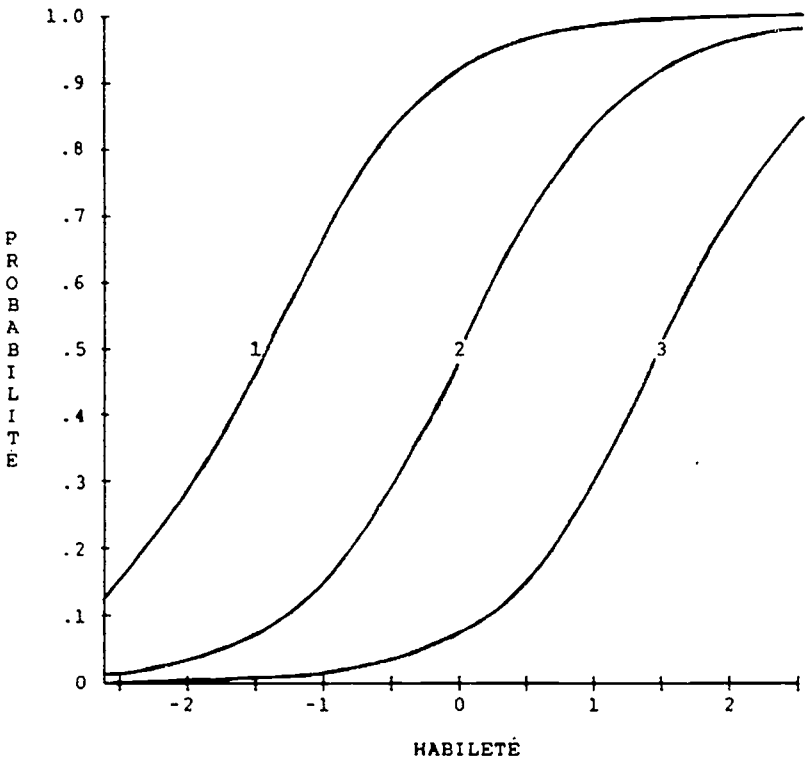
Hambleton et Swaminathan (1985:chap 7) décrivent cinq procédures pour effectuer l'estimation des paramètres:

- le maximum de vraisemblance combiné;
- le maximum de vraisemblance conditionnel (avec le modèle de Rasch);
- le maximum de vraisemblance marginal (avec le modèle de Rasch ou le modèle à deux paramètres);

- l'estimation bayésienne;
- l'estimation par approximation (avec les modèles à deux ou trois paramètres).

Des recherches sont encore en cours afin de rendre ces procédure plus efficaces. Les travaux autour des procédures bayésiennes, où l'on amorce la calibration à partir d'hypothèses sur la distribution des réponses, sont particulièrement prometteurs (Mislevy 1986).

FIGURE 2.7
Courbes d'un modèle à un paramètre



1 : b = -1.5
2 : b = 0
3 : b = 1.5

Les estimations qui résultent de l'application de ces procédures ont l'avantage d'être indépendantes du groupe de sujets à partir duquel on a obtenu les données. Ainsi, contrairement aux indices fournis par la théorie classique, les paramétres ne sont pas affectés par la distribution des réponses à l'intérieur du groupe. Une fois la calibration complétée, on peut, à l'aide de procédures semblables à celles qu'on utilise pour l'estimation des paramètres, faire l'estimation de l'habileté des sujets. De même qu'il y a invariance des estimations relatives aux items, il y a invariance des estimations de l'habileté des sujets. Cette propriété des estimations de l'habileté des sujets est déterminante dans toutes les applications qui font appel à une banque d'items. En effet, elle implique qu'une fois les items calibrés, on peut estimer l'habileté des sujets même si ceux-ci ont répondu à des items différents.

2.2.2 *Les contraintes*

Outre l'invariance des paramètres des items et l'invariance des estimations d'habileté, Bejar (1983) ajoute comme avantage de la théorie du trait latent, par rapport à la théorie classique, le fait qu'on puisse obtenir une indication de la précision de la mesure en fonction de l'habileté. Il ajoute toutefois que tous ces avantages ne tiennent qu'à la condition que les exigences de la théorie soient respectées. Ces exigences sont de trois ordres: indépendance locale, unidimensionalité et adéquation du modèle choisi. Pour McLean et Ragsdale (1983) ces exigences sont si fortes que les modèles issus de la théorie ne conviennent pas du tout aux situations réelles où un test est nécessaire.

2.2.2.1 *L'indépendance locale*

Le postulat d'indépendance locale implique qu'à un niveau d'habileté donné, les réponses des sujets sont statistiquement indépendantes. En d'autre termes, lorsque Θ est constant, on ne doit pas retrouver de corrélation entre chacun des items. Ainsi, si un groupe d'items se distingue d'un autre groupe d'items par le contenu

particulier qu'il vérifie, on observera des corrélations entre les items à l'intérieur des groupes et le postulat d'indépendance locale ne sera pas respecté. Selon Traub (1983:61), il s'agit là d'une exigence peu réaliste de par la nature même de l'apprentissage: *It seems reasonable then to expect differences of many kinds, some obvious, some subtle, in what it is different students learn, both in school and outside.* L'exigence est d'autant plus restrictive, dans le cas de tests en langue seconde, que celle-ci est apprise dans des conditions qui peuvent varier considérablement d'un apprenant à l'autre.

L'indépendance locale implique aussi que la performance à un item n'influence pas la performance à d'autres items. S'il s'avérait par exemple que les tests de closure fassent intervenir plus que le contexte immédiat et que de ce fait l'identification d'un mot, fournisse un indice pour l'identification du mot suivant ou du mot précédent, il faudrait alors renoncer à en calibrer les items. De fait, le problème risque de se poser pour toute mesure intégrative où l'on reconnaît une interaction entre les éléments vérifiés.

2.2.2.2 *L'unidimensionalité*

Nous avons déjà souligné l'importance du concept d'unidimensionalité quand on veut appliquer la théorie du trait latent. Cette question est si importante que c'est sans doute l'aspect de la théorie qui a suscité le plus de controverses et de discussions. Pourtant, il importe de préciser que l'unidimensionalité est étroitement liée à l'indépendance locale des items. L'observation de la matrice des corrélations entre les items de tests multidimensionnels montre en effet que des réseaux d'items se constituent autour des dimensions du test.

Hattie (1981, 1985) décrit un grand nombre de techniques pour détecter la multidimensionalité sans toutefois en trouver une qui soit tout à fait appropriée. Il distingue quatre familles de techniques:

— *Les échelles de Guttman:*

On organise les données de façon à ce que la diagonale de la matrice des réponses sépare les réponses exactes de celle qui sont inexactes. Soit un test de n items administré à n + 1 sujets, formant une échelle parfaite; si n = 5, 1 indique une réponse correcte et 0 une réponse incorrecte, on obtient:

1	1	1	1	1
1	1	1	1	0
1	1	1	0	0
1	1	0	0	0
1	0	0	0	0
0	0	0	0	0

La construction d'une échelle implicationnelle de ce type suppose l'unidimensionalité. Cependant, il est peu vraisemblable que les données d'un test puissent se conformer à une échelle aussi contraignante.

— *Les indices de fiabilité:*

Des indices comme le KR-20 ou l'*alpha* de Cronbach peuvent refléter la structure dimensionnelle d'un test mais sont davantage des indices de consistance interne que d'unidimensionalité. Ainsi, à moins de calculer le coefficient de fiabilité en comparant la première et la deuxième moitié du test, les indices de fiabilité peuvent être assez élevés dans le cas de test où la contrainte de temps est importante. Or, dans ces cas, il y a certainement une dimension, la vitesse, qui se superpose à l'habileté sur laquelle porte l'ensemble du test. De plus, on sait que les indices de fiabilité varient avec le nombre d'items de sorte qu'il est difficile d'établir des références précises. Par ailleurs, il faut noter que les indices de discrimination, particulièrement la corrélation bisérielle qui sert dans l'analyse des items pour améliorer la fiabilité, peuvent aussi servir à repérer des items dont le contenu s'écarte d'une dimension commune.

— *La décomposition en facteurs:*

L'analyse factorielle des corrélations entre les items pose des problèmes particuliers. En utilisant les corrélations phi, on obtient

rarement des résultats satisfaisants du fait de l'émergence d'un facteur de difficulté. Hulin, Drasgow et Parsons (1983:chap 8) suggèrent d'analyser plutôt les corrélations tétrachoriques. Selon Lord et Novick (1968:349,382) l'analyse des corrélations tétrachoriques, quand elle réussit, tend à favoriser une interprétation unidimensionnelle. McDonald (1980) propose une analyse non linéaire qui tiendrait mieux compte du fait que la théorie du trait latent reconnaît la non linéarité de la relation entre la performance à un item et l'habileté. Reckcase (1978) examine les *eigenvalues* produites par l'analyse de matrices de corrélations tétrachoriques et conclut en la valeur de l'analyse factorielle pour ce type de problèmes. De plus, il note que la robustesse des procédures d'estimation des paramètres autorise l'utilisation de la théorie du trait latent même lorsque les données ne sont pas parfaitement unidimensionnelles. À l'aide d'une procédure similaire, Davidson (1988) analyse plusieurs tests d'anglais langue seconde et remarque que la grande majorité de ces tests sont unidimensionnels. Face aux problèmes associés aux corrélations entre items, Cook *et al.* (1988) proposent de regrouper les items par série de 3 à 7 items de même type et de difficulté égale. En appliquant la technique à la section sur l'aptitude verbale du *Scholastic Aptitude Test* (*SAT*), ils remarquent que le test est plutôt unidimensionnel bien que le sous-test de lecture se distingue par la présence d'un facteur supplémentaire qui pourrait bien être attribuable à des contraintes de temps.

— *Les analyses du trait latent:*

Dorans et Kingston (1985) complètent l'analyse factorielle par une analyse du trait latent de sections portant sur l'aptitude verbale d'un test du même type que le *SAT*, le *Graduate Record Examination* (*GRE*). Il dégagent également deux facteurs fortement corrélés: un facteur de lecture et une maîtrise des éléments discrets. La technique utilisée par Dorans et Kingston s'apparente à la procédure de Bejar (1980) qui propose aussi, comme alternative aux techniques d'analyse factorielle, de diviser le test entier en fonction des différences de contenu qu'on y trouve de façon à composer plusieurs sections. Il s'agit alors de comparer les résultats de la calibration pour le test complet avec les résultats des calibrations par section.

La procédure permet d'identifier les tests multidimensionnels à la condition qu'aucun facteur principal ne se dégage clairement. En effet, Harrison (1986:107) observe que des programmes d'estimation des paramètres comme *LOGIST* (Wingersky *et al.* 1982), fournissent des estimations relativement robustes: *As a single group factor controls variation in more items and concomitantly in a large percentage of items composing a test, LOGIST begins to take this factor into account as part of the unidimensional trait.* Hambleton et Rovinelli (1986) donnent des exemples de cas de multidimensionalité qui échappent à la procédure de Bejar. Henning (1988), quant à lui, juge la procédure assez efficace après avoir comparé les calibrations d'un test multidimensionnel et d'un test unidimensionnel: il note que même si les paramètres de difficulté varient peu, les estimations des niveaux d'habileté des sujets divergent sensiblement.

Par ailleurs, avec les modèles à deux ou trois paramètres, il est possible d'utiliser l'indice *b* afin de repérer les items qui ne se conforment pas à la dimension commune pour éventuellement les retrancher et recalibrer. On obtient alors un test plus unidimensionnel et des coefficients d'adéquation du modèle plus satisfaisants. Si l'élimination des items divergents ne suffit pas, on pourra alors constituer des sous-tests pour chacune des dimensions ou recourir à des procédures plus complexes bien qu'encore imparfaites pour traiter des données multidimensionnelles. Il faut également noter que la plupart des logiciels de calibration fournissent des indices sur l'adéquation du modèle par rapport aux données. McNamarra (1990) propose même d'utiliser ces indices à des fins de validation pour les tests de langue seconde.

2.2.2.3 *L'adéquation du modèle*

Les indices d'adéquation que calculent les logiciels de calibration permettent de repérer les ensembles de réponses qui ne se conforment pas au modèle choisi. Il peut s'agir d'items pour lesquels, à un niveau d'habileté donné, les réponses présentent des écarts importants d'un sujet à l'autre ou dont la pro-

portion de réponses correctes s'écarte de la probabilité prévue par la courbe caractéristique de l'item. Il peut aussi s'agir de sujets dont les configurations de réponses n'obéissent pas à ce que prédit le modèle. Les sources des divergences des items ou des sujets sont généralement reliées à un problème d'unidimensionalité ou à un mauvais choix de modèle. Au plan de l'adéquation des sujets au modèle, Traub (1983:64) fait remarquer qu'il est peu réaliste, voire dangereux, de chercher à faire correspondre les comportements des apprenants à des modèles aussi rigides:

> It will be a sad day indeed when our conception of measurable educational achievement narrows to the point where it coincides with the criterion of fit to a unidimensional item response model, regardless of which model is being fitted.

Au plan de l'adéquation des items, Traub signale qu'il est certainement abusif de croire que tous les items d'un test puissent discriminer de la même façon. Aussi préfère-t-il au modèle de Rasch, un modèle à deux paramètres ou, dans le cas de questions à choix multiple, un modèle à trois paramètres.

Néanmoins, l'avantage du modèle à trois paramètres est souvent remis en cause de par la taille de l'échantillon qu'il impose. Si on peut obtenir des estimations raisonnables avec un modèle à un paramètre en utilisant 200 sujets, il en faut souvent dix fois plus pour en arriver à un degré de précision comparable avec un modèle à trois paramètres. Ainsi que le fait remarquer Lord (1983): *Small N justifies Rasch model.*

Traub et Lam (1985) font remarquer que l'augmentation du nombre de cas ne garantit cependant pas une meilleure adéquation et en viennent à douter de la valeur de la théorie du trait latent. Il n'en reste pas moins que la théorie compte de plus en plus d'adeptes et que, dans l'évaluation de la langue seconde, on voit s'implanter le modèle de Rasch. Henning *et al.* (1985:152) font remarquer:

Item Response Theory in general and Rasch in particular, are sufficiently robust with regard to the assumption of unidimensionality to permit applications to the development and analysis of language tests which may be comprised of item domain representing diverse subskills of language use and which may be applied in the testing of persons from diverse national, linguistic, cultural, educational, and professionnal backgrounds.

2.2.3 *La fonction d'information*

Dans les cas où l'on peut satisfaire les exigences de la théorie du trait latent, celle-ci devient fort séduisante surtout à cause de la notion d'information qui est sous-jacente à la plupart des applications de la théorie.

Dans le cadre de la théorie classique, on suppose que le nombre de bonnes réponses à un test est l'indication la plus juste qu'on peut obtenir sur le score réel du sujet c'est-à-dire la performance de celui-ci indépendamment de l'erreur inhérente à l'instrument de mesure. Dans le cadre de la théorie du trait latent, cette assertion ne tient que si on utilise un modèle à un paramètre, où tous les items discriminent également et où le hasard ne joue pas. On dira alors que pour le modèle de Rasch, le nombre de réponses exactes à un test de longueur pré-établie, est une statistique suffisante, c'est-à-dire un indice qui tient compte de toute l'information disponible. Ainsi:

$$\theta = \sum_{i=1}^{n} U_i \qquad (2.4)$$

où pour un test de n items i, U prend la valeur de 0 pour une réponse inexacte et de 1 pour une réponse exacte. Avec un modèle à deux paramètres, il faut tenir compte de la discrimination en appliquant la formule 2.5:

$$\theta = \sum_{i=1}^{n} aU_i \qquad (2.5)$$

Cependant, avec un modèle à trois paramètres, même si le nombre de réponses exactes peut parfois représenter une ap-

proximation acceptable (Yen 1984), une telle formule n'existe pas. De même, si le nombre d'items varie d'une administration à l'autre, il faut recourir à d'autres moyens pour estimer l'habileté. Une fois les paramètres connus, on peut utiliser une procédure par maximum de vraisemblance qui consiste à résoudre l'équation 2.6:

$$\sum_{i=1}^{n} (U_i - P_i) \frac{P'_i}{P_i Q_i} = 0 \qquad (2.6)$$

où $Q_i = 1 - P_i$ et P'_i représente la première dérivée:

$$P'_i = dP_i / d\theta.$$

Non seulement le maximum de vraisemblance fournit une statistique suffisante, mais il assure aussi la normalité asymptotique de sorte que la moyenne de l'estimation s'établit à Θ et l'écart-type à $[I(\Theta)]^{-1}$. Cette dernière valeur constitue en fait l'erreur type de la mesure pour un sujet donné.

L'erreur de la mesure est inversement reliée à la quantité d'information que fournit le test: plus le test apporte d'information sur le sujet, moins il y a de possibilité d'erreur. Chaque item contribue à minimiser l'erreur. La fonction d'information s'écrit donc ainsi:

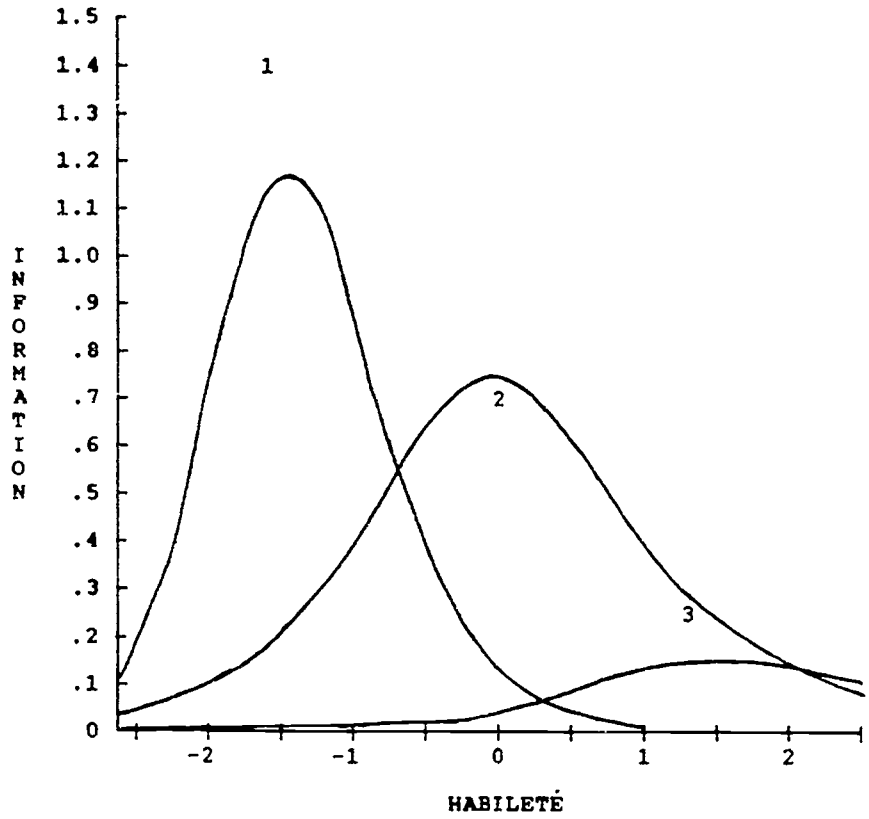
$$I(\theta, U_i) = \frac{P'_i}{P_i Q_i} \qquad (2.7)$$

La figure 2.8 montre quelques exemples de courbes d'information d'items. L'information obtenue varie en fonction des paramètres et de l'habileté. Le sommet de la courbe correspond au niveau d'habileté où l'item est le plus efficace.

La fonction est d'autant plus intéressante que l'information est cumulative. Ainsi,

$$I(\theta) = \sum_{i=1}^{n} I(\theta, U_i) \qquad (2.8)$$

FIGURE 2.8
Courbes d'information d'items



1 : a = 1.4 b = -1.5 c = 0.1
2 : a = 1.0 b = 0 c = 0
3 : a = 0.6 b = 1.5 c = 0.3

Pour chaque item qui s'ajoute au test, on peut prévoir dans quelle mesure cet item contribuera à l'information totale. Cette propriété d'additivité permet de déterminer les zones d'habileté où le test est le plus efficace. La figure 2.9 montre la courbe d'information de deux tests typiques comprenant chacun une vingtaine d'items. Le premier est plutôt facile mais comprend des items qui s'adressent à des sujets répartis sur une large gamme de niveaux; le second, plus

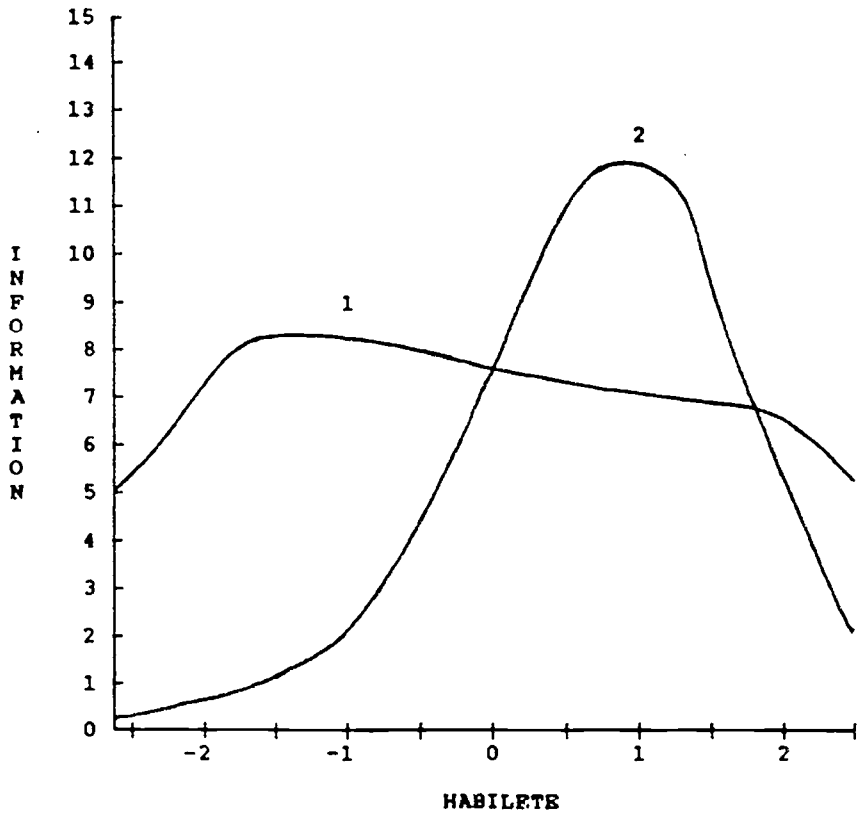
difficile, vise à recueillir davantage d'information autour d'un niveau particulier. Le premier pourrait bien être un test de classement alors que le second servirait plutôt à la sélection de candidats.

À partir de la fonction d'information, on peut calculer l'erreur type du test pour chaque niveau d'habileté:

$$E(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

$$[I(\theta)]^{-\frac{1}{2}} \qquad (2.9)$$

FIGURE 2.9
Courbes d'information de tests



1: Test de classement
2: Test de sélection

La fonction d'information présente un avantage incontestable par rapport à la notion de fiabilité de la théorie classique. Alors que la notion de fiabilité suppose que la marge d'erreur est identique, peu importe le niveau d'habileté, la fonction d'information permet d'identifier des zones où l'erreur sera plus ou moins grande. Cette fonction utilise les paramètres fournis par la calibration et est le fondement des applications de la théorie du trait latent.

2.2.4 *Les applications*

Selon Lord (1977c), les applications de la courbe caractéristique des items et de la courbe d'information qui en découle, montrent bien que, contrairement à la théorie classique, la théorie du trait latent permet de faire des prédictions. De ce point de vue, la théorie s'apparente à la théorie de la généralisabilité (Brennan 1983, Linn et Werts 1979 et, pour les tests en langue seconde, Bolus *et al.* 1982). Quatre types d'applications nous semblent particulièrement intéressants.

2.2.4.1 *L'élaboration de tests critériés*

La propriété d'invariance des items permet de dépasser l'évaluation normative et de situer les sujets sur une échelle d'habileté indépendante de la population qui sert à la standardisation du test. Une fois la validité de la mesure assurée, on obtient donc un résultat qui peut être directement relié à des objectifs d'acquisition. D'autre part, l'invariance des sujets permet de comparer des candidats sans que ceux-ci n'aient nécessairement répondu aux mêmes questions. On peut alors choisir les items qui apportent le plus d'information, compte tenu des objectifs du test et de la zone d'habileté relativement à laquelle des décisions devront être prises. Les modèles de réponses aux items deviennent alors des outils précieux pour élaborer des tests critériés qui peuvent s'adapter à des besoins particuliers (Yen 1984). On

peut même entrevoir des utilisations avec des tests à réponses ouvertes qui ne sont pas nécessairement corrigés dichotomiquement (Samejima 1978).

2.2.4.2 *L'équivalence entre les versions d'un test*

Afin de préserver la confidentialité d'un test ou d'éliminer l'effet de répétition, il est souvent utile de disposer de versions comparables. L'examen des courbes d'information des tests offre une alternative aux techniques habituelles d'équivalence (Angoff 1971, 1982, Morris 1982, Samejima 1977). On peut en effet obtenir des versions équivalentes d'un test composé à partir d'une banque d'items en s'assurant d'une part, qu'on mesure un contenu similaire unidimensionnel et d'autre part, que les courbes d'information des différentes versions soient identiques. Ainsi, depuis 1978, on utilise la théorie du trait latent pour établir l'équivalence entre les différentes versions du *TOEFL* (Cowell 1982).

2.2.4.3 *La détection de biais*

Si les résultats de la calibration d'une série d'items à partir des réponses d'un groupe de sujets divergent des résultats de la calibration des mêmes items à partir des réponses d'un groupe de sujets culturellement différent du premier, on peut conclure que le test favorise un groupe culturel par rapport à l'autre (Vetterli 1987). En effet, bien que le principe d'invariance des items n'impose pas une distribution normale de l'habileté dans le groupe servant à la calibration, la composition du groupe doit cependant respecter le postulat d'unidimensionalité. Madsen (1986) démontre que puisque l'analyse est affectée par la multidimensionalité des sujets, on peut détecter les biais culturels d'un test de langue seconde. Par ailleurs, de par la valeur prédictive des courbes caractéristiques des items, on peut repérer des sujets qui présentent une configuration de réponses aberrante (Hulin *et al.* 1983:chap 4-5, Levine et Drasgow 1983). On doit alors se demander si le sujet veut déjouer le test, s'il a fait ce test auparavant, s'il s'agit d'une variation individuelle au plan des connaissances ou s'il s'agit d'un biais culturel.

2.2.4.4 *Le testing adaptatif*

Lorsque les items d'une banque ont été calibrés, il est possible d'ajuster le niveau de difficulté après chaque réponse et de poursuivre l'administration jusqu'à ce qu'on ait atteint le niveau d'information souhaité. Urry (1977) précise certains critères pour les paramètres d'un test. Selon lui, on peut accroître rapidement la fonction d'information si la valeur des paramètres de difficulté *a* est distribuée uniformément à l'intérieur de la gamme d'habilité visée, si la valeur des paramètres de discrimination *b* dépasse .8 et si on maintient la valeur des paramètres de hasard *c* à un niveau assez bas. Ces critères devraient servir de lignes directrices pour l'élaboration d'un bon test adaptatif. Si le test s'avère valide, on obtient alors un test sur mesure (*tailored test*) qui, comme le fait remarquer Séguin-Duquette (1982) en se référant à l'évaluation de la langue seconde, s'inscrit bien dans le courant de l'enseignement individualisé.

2.3 **Un test adaptatif en langue seconde**

Lorsqu'on envisage d'utiliser un test adaptatif en langue seconde, on peut, soit recourir à un instrument déjà existant, soit en élaborer un. Comme il existe pour l'instant peu de tests adaptatifs standardisés et que les besoins des établissements varient considérablement, la dernière solution peut présenter certains avantages même si elle implique un investissement de temps, d'énergie et d'argent assez important. De fait, au moment où nous avons proposé le présent projet de recherche, il n'existait, à notre connaissance, aucun didacticiel de testing adaptatif de langue seconde qui fasse usage de données réelles. Le fait de mener une expérimentation avec des items et des sujets réels plutôt que simulés s'avérait donc une tâche considérable mais présentait un intérêt certain (Tung, communication personnelle).

2.3.1 ***Les tests adaptatifs disponibles***

Les seuls tests adaptatifs qui aient été mis sur le marché sont le produit d'une équipe de l'Université Brigham Young. Par

ailleurs, des tests intéressants conçus par des groupes de travail oeuvrant d'abord dans le domaine de l'évaluation de l'expression orale, devraient bientôt voir le jour. La liste des tests adaptatifs disponibles ou en voie de réalisation est donc relativement courte.

2.3.1.1 *Le TOEFL informatisé*

Bien que le test n'ait pas été utilisé avec des sujets réels (Hicks 1986) et qu'il n'exploite pas pleinement les possibilités du testing adaptatif, il présente un intérêt du fait qu'il s'agit de la version informatisée du test standardisé en langue seconde le plus populaire au monde. Pour la version informatisée, on a retenu 19 items de la section II du *TOEFL* (Structure et expression écrite) et 28 items de la section III (Lecture et vocabulaire). Ces items sont distribués dans cinq niveaux: A, B, C, D et E. L'administration commence avec un item de niveau moyen. Le test s'arrête quand le sujet a répondu correctement à quatre questions appartenant à un niveau. Le sujet reçoit des items de trois niveaux adjacents de sorte qu'il est classé dans une des trois catégories suivantes: ABC, BCD ou CDE. Le test apparaît comme un moyen efficace de faire un premier tri parmi un groupe d'étudiants.

2.3.1.2 *Le S-CAPE*

Ce test vise à classer les étudiants qui s'inscrivent à des cours d'espagnol à l'Université Brigham Young (Larson 1987). Il utilise une banque d'un millier d'items à choix multiple évaluant la lecture ou portant sur la grammaire et le vocabulaire. En utilisant le modèle de Rasch, on a pu limiter à moins de 200 le nombre de sujets nécessaires pour la calibration. Les items sont distribués dans une cinquantaine de niveaux d'habileté et ont été calibrés selon le modèle de Rasch. L'épreuve commence par un item facile et se termine quand l'étudiant a répondu correctement à cinq questions d'un certain niveau ou incorrectement à quatre questions d'un certain niveau.

2.3.1.3 *Le CompuTest*

Fruit de trois années de travail (Madsen 1989a, 1989b), il s'agit d'un test de classement en anglais langue seconde. La version la plus récente utilise un millier d'items calibrés selon le modèle de Rasch. Le critère d'arrêt est le résultat de la fonction d'information. L'organisation du contenu rappelle les premières versions du *TOEFL* (Oller et Spolsky 1978). On y retrouve quatre sections: écoute, lecture, vocabulaire, et grammaire. Il faut noter que comme on évalue la compréhension auditive, on doit coupler un micro-ordinateur de la famille IBM à un cassettophone ou à un appareil *CD-Rom*.

2.3.1.4 *Le test de l'ACTFL*

L'élaboration de ce test de lecture s'insère dans un effort pour mettre au point une série de tests destinés à mesurer les habiletés réceptives en suivant les lignes directrice établies par l'ACTFL (Dandonelli 1987, Dandonelli et Rumizen 1989). Les textes à lire figurent dans un questionnaire alors que les questions apparaissent à l'écran. Le test dure environ une demi-heure. L'administration fait appel à un sous-programme du logiciel *MicroCAT* conçu pour la famille IBM (Assessment System Corp. 1987). On s'est aussi servi de ce logiciel pour faire la calibration des items selon un modèle à deux paramètres. L'originalité de ce test tient au fait qu'on postule que la lecture est une activité multidimensionnelle (Kaya-Carton et Carton 1986) et qu'on doit donc recourir à des techniques appropriées. Cela implique une analyse beaucoup plus complexe et une expérimentation à grande échelle: on a administré 750 items à 2,500 sujets.

2.3.1.5 *Le test de la Défense américaine (DLI)*

Ce vaste projet vise à mettre au point une série de tests adaptatifs pour mesurer la lecture dans plusieurs langues (Lowe et Jarczewski 1989). Le premier prototype est en hollandais. Comme le test doit servir à des décisions importantes, on administre de 100 à 150 items par session. La plupart se présentent comme des

questions de compréhension sur un passage à lire, mais on a aussi ajouté des items à élément discret. La calibration se fait selon le modèle de Rasch. Ce test se distingue par une procédure complexe de sélection des items. Premièrement, le sujet est exclu d'un niveau s'il échoue à quatre questions de ce niveau. Deuxièmement, on essaie de respecter le déroulement typique de l'entrevue orale proposée par L'*InterAgency Language Roundtable*: mise en train, progression, épreuves de niveaux et remise à niveau. Troisièmement, la sélection prend en considération les aspects culturels et le contenu des textes présentés.

On pourrait ajouter à cette liste la version informatisée du test d'anglais langue étrangère de Oxford (Willmot et Kam Chuan Aik 1990), mais dans ce cas, la possibilité de générer différentes épreuves sert plutôt à assurer la sécurité du test qu'à l'adapter à l'étudiant. De même, il faut noter des applications de la théorie du trait latent telles que les travaux de Griffin (1985) sur l'entrevue orale, de Zettersten (1985) sur les connaissances lexicales ou de De Jong (1986) pour la construction de test par niveaux. Pourtant, il ne s'agit pas de tests adaptatifs à proprement parler. On constate donc que les tests adaptatifs sont peu nombreux et souvent encore en développement.

2.3.2 *La création d'un test adaptatif*

En l'absence de test adaptatif pour le classement général en français langue seconde, l'élaboration d'un tel instrument constitue un aspect majeur de la contribution de la présente recherche. Nous avons suivi les trois étapes prescrites par Henrysson (1971) pour la mise au point d'un test: la pré-expérimentation, l'expérimentation et l'administration expérimentale. Toutefois, il va sans dire que la mise au point d'un test adaptatif est plus complexe que celle d'un test conventionnel.

2.3.2.1 *La planification*

Au plan pédagogique, il a d'abord fallu choisir un cadre théorique dont nous avons précisé les grandes lignes dans les pages

précédentes. Nous en sommes alors venu à un test de classement en trois parties: compréhension, choix de l'énoncé approprié et phrases à trou. Il est clair cependant que cette forme de test tient non seulement compte d'orientations pédagogiques mais qu'elle doit concilier deux modes d'administration fort différents. D'une part, le test «papier-crayon», sans bande audio avec grille de correction. D'autre part, un test informatisé qui, malgré un progrès technologique rapide, doit respecter les limites des micro-ordinateurs qu'on trouve aujourd'hui dans les établissements. De plus, il nous a fallu choisir un cadre théorique au plan docimologique. Endossant les conclusions de Henning (1984), nous pensons que la théorie du trait latent, dans la perspective de notre recherche, convient bien à l'élaboration d'un test adaptatif en langue seconde. Malgré les réserves dont l'invariance des sujets et l'invariance des items ont fait l'objet, la souplesse que permettent ces principes explique pourquoi la plupart des tests adaptatifs qu'on connaît ont été élaborés dans le cadre de la théorie du trait latent.

La première étape dans la création d'un test adaptatif consiste à rédiger un grand nombre d'items qui pourront être intégrés dans un questionnaire. Tous les items sont administrés séquentiellement, soit de façon conventionnelle, soit par ordinateur. En apportant un grand soin à la rédaction des items, on peut éviter que par la suite l'analyse élimine tellement d'items qu'il faille reprendre l'opération. Celà est d'autant plus important que des facteurs d'ordre pratique et psychologique limitent le nombre d'items que l'on peut administrer en une session. Même si, par la suite, les techniques d'ancrage permettent d'élargir la banque, il est souhaitable qu'on puisse après la première calibration disposer d'une quarantaine de bons items dans chaque banque. On détermine les items à conserver tant à partir des indices classiques que des indices que fournit la calibration.

2.3.2.2 *La calibration*

À cause de sa simplicité et surtout parce qu'il peut fonctionner avec des échantillons plus restreints, le modèle de Rasch connaît

une certaine popularité pour ce qui est de la paramétrisation des items (Auger 1986). Plusieurs logiciels sont disponibles pour effectuer les longs calculs qu'impliquent les procédures itératives de la calibration: par exemple, *BICAL* (Wright *et al.* 1979), MicroScale (Madsen 1989) ou le sous-programme *RASCAL* de *MicroCAT* (Assessment System Corp 1987). Compte tenu de la taille de l'échantillon de la pré-expérimentation, nous avons utilisé tout d'abord le modèle de Rasch.

Toutefois, il nous semblait qu'un modèle à trois paramètres était plus approprié. En effet, nous n'avions aucune raison de penser que les items discriminaient tous de la même manière pas plus que nous n'avions de raison de penser que, dans un test à choix multiple, l'effet de hasard était négligeable. Ree (1981:18) indique que même si un échantillon de 2,000 sujets permet de minimiser la marge d'erreur avec un modèle à trois paramètres, des échantillons plus modestes peuvent suffire: *If an ordering of examinees is all that is required or if the relatively higher errors are not important to the purpose, item polls of 100 items calibrated on a sample of 500 subjects will produce high correlations, especialy if 20 or more items are administered.* Un échantillon de 750 sujets représente donc un objectif réaliste et acceptable pour la première version d'un test de classement. Les paramètres peuvent se préciser en intégrant les sujets qui utilisent par la suite les versions conventionnelles standardisées ou les résultats aux items d'ancrage. La calibration des items avec un modèle à deux ou trois paramètres peut se faire avec des logiciels comme *LOGIST* (Wingersky *et al.* 1982), *BILOG* (Mislevy et Bock 1986), *MultiLOG* (Thissen 1986) ou le sous-programme *ASCAL* de *MicroCAT* (Assessment System Corp 1987).

Il faut préciser également que la calibration n'exclut pas le recours aux techniques classiques d'analyse des items. Au plan du test, l'examen des moyennes, des variances, des indices de fiabilité a encore sa place. Au plan des items, il est souvent utile d'étudier les corrélations bisérielles ou point-bisérielles et les indices de probabilité, tant pour les bonnes réponses que pour les distracteurs. Des logiciels comme *LERTAP* (Nelson 1970), le sous-programme *ITEMAN*

de *MicroCAT*, ou plusieurs sous-programmes de logiciels de statistiques plus générales peuvent être utilisés pour réaliser une analyse des items selon les principes de l'analyse classique.

2.3.2.3 *La programmation*

Si les tests adaptatifs sont encore peu nombreux, les logiciels qui servent à la programmation de tels tests le sont aussi. Même des programmes comme *CALIS* (Duke University 1989), conçus essentiellement pour l'enseignement de la langue et dotés de fonctions pour l'administration de tests, se prêtent mal à la manipulation d'items rangés dans des banques et ont des capacités de calculs insuffisantes. Le logiciel *MicroCAT* contient une série de sous-programmes très puissants qui peuvent servir à développer et administrer un test adaptatif. On peut y intégrer des graphiques, relier divers sous-tests ou choisir divers algorithmes de sélection. Toutefois, l'administration à plusieurs étudiants suppose qu'on fasse l'acquisition de plusieurs systèmes d'administration ce qui rend les coûts prohibitifs. Certains compilateurs conçus spécifiquement pour gérer des banques de données peuvent être utilisés. L'utilisation d'un langage de programmation demeure une alternative dont les possibilités sont infinies mais qui requiert beaucoup de temps et une certaine formation. Nous avons, quant à nous, utilisé le langage *Turbo-Pascal* complété de quelques fonctions pré-définies pour la gestion de base de données (Borland 1985, Borland 1987).

Lors de la mise en place d'un système de test adaptatif, il faudra prévoir la programmation de deux composantes essentielles:

— *Une composante de développement:*

C'est le système de gestion de la (des) banque(s). C'est grâce à ce système qu'on peut corriger certains items, en ajouter ou en éliminer. Il est également souhaitable que la composante de développement puisse servir à simuler l'administration de séances de testing adaptatif.

— *Une composante d'administration:*

C'est le système qui sert à calculer le niveau, à choisir les items et à les présenter au sujet. La composante d'administration sert également à rapporter les résultats, transformer les scores, informer l'étudiant, enregistrer le niveau dans un fichier...

On peut envisager diverses procédures pour l'estimation du niveau. Nous avons opté pour une estimation basée sur le maximum de vraisemblance avec une procédure alternative au début du test et dans les cas de non-convergence. Par ailleurs, dans le cas de notre test, il apparaissait *a priori* plus raisonnable de traiter chaque section comme un sous-test indépendant. Le résultat du sous-test précédent peut servir de point de départ au sous-test suivant. Pour le premier sous-test, on utilise des renseignements que fournit le sujet au tout début. Le fait de disposer d'une évaluation préliminaire permet de réduire le nombre d'items nécessaires pour atteindre une marge d'erreur acceptable.

③

L'ÉLABORATION DU TEST «PAPIER-CRAYON»

3.1 De la version pré-expérimentale à la version expérimentale

Prévoyant un taux de rejet des items entre 30% et 40% et conscient que la durée moyenne du test ne devait pas dépasser deux heures, cinquante items par section nous semblaient raisonnables. En éliminant un item sur trois et en les remplaçant après la première mise à l'essai, on pouvait espérer obtenir au moins 40 items dans la banque. Compte tenu des objectifs de la recherche, il nous semblait prudent de nous restreindre à des questions à choix multiple. Par ailleurs, afin d'assurer une bonne fiabilité sans prolonger indûment l'administration du test, nous avons décidé de présenter quatre choix par item. Même si les distracteurs font référence à des erreurs susceptibles d'être commises par des sujets anglophones, on ne cherchait jamais à «piéger» l'étudiant.

3.1.1 *L'expérimentation*

3.1.1.1 *La rédaction des items*

Dans la mesure où l'exploitation d'une banque d'items se prête mal à une sélection en terme de contenu spécifique et que le test n'a pas de fonction diagnostique, il nous semblait assez peu utile de procéder à un inventaire rigoureux des aires de contenu devant être représentées. Ainsi, nous n'avons pas fait de liste exhaustive et structurée des points de langue à vérifier ou des

situations à illustrer. Nous avons simplement veillé à éviter les redondances superflues; nous nous sommes assuré de vérifier les points de langue les plus importants et de faire référence à des situations relativement familières. En mettant à contribution notre riche expérience dans l'enseignement du français langue seconde, il nous était possible de prévoir les difficultés propres à chaque niveau et de créer des items appropriés. Par ailleurs, soucieux de préserver le caractère intégratif de l'habileté à mesurer, nous n'avons pas cherché à isoler l'élément vérifié de difficultés susceptibles d'apparaître concurremment.

Dans cette perspective, le premier sous-test évalue la compréhension globale plutôt que des éléments grammaticaux ou lexicaux spécifiques. Pourtant, on remarque une insistance sur la compréhension des relations temporelles (la chronologie) ou logiques (la cause par rapport à l'effet). Les textes à lire contiennent environ 35 mots. Comme cette partie implique plus de lecture dans la langue seconde, elle a été placée au début du test. On demande à l'étudiant de reformuler le contenu ou de répondre à une question. Ainsi, à l'item 47, l'étudiant doit reformuler en résumant le contenu d'une carte postale:

> Bonjour Pierre! Je passe des vacances magnifiques. Je viens d'arriver à Marseilles. C'est une ville très spéciale. J'ai hâte de me retrouver sur les plages de la Côte d'Azur et de pouvoir me baigner dans la Méditerranée. À bientôt! Jacques.
>
> a) Jacques est en vacances en Italie.
> b) Jacques passe de très bonnes vacances.
> c) Jacques va passer les prochains jours à Paris.
> d) Jacques revient chez lui dans quelques jours.

S'il doit répondre à une question, la question peut être directe, comme dans le cas de l'item 38:

> Comme il faisait gris, je n'ai pas mangé sur la terrasse. L'après-midi, je suis allé à la bibliothèque. Quand je suis sorti, il pleuvait. J'ai décidé de prendre le taxi pour rentrer chez moi.

Quand est-ce que la pluie a commencé?
a) Avant qu'il mange.
b) Pendant qu'il était à la bibliothèque.
c) Pendant qu'il était dans le taxi.
d) Après qu'il est arrivé chez lui.

Ailleurs, comme à l'item 45, la question prend la forme d'un énoncé à continuer:

Nous vous prions de prendre note que la partie de baseball prévue pour cet après-midi est annulée à cause de la grève des joueurs. Nous ne pouvons malheureusement rembourser aucun billet.

La partie est annulée...
a) parce qu'il pleut.
b) parce que les joueurs ont arrêté de travailler.
c) parce que l'équipe a des problèmes financiers.
d) parce que plusieurs joueurs sont malades.

Dans le deuxième sous-test, nous avons tenté de présenter des situations familières. Dans plusieurs cas, l'étudiant doit choisir l'énoncé qui est le plus approprié sémantiquement. Ainsi l'item 16, conçu à l'intention des débutants, se lit ainsi:

You are driving too fast and a policeman asks you to stop on the shoulder. What do you expect the policeman to say?

a) Votre permis de conduire s'il vous plaît.
b) Haut-les-mains!
c) Je m'appelle Jean-Marc Labonté.
d) Le plein, s'il vous plaît.

Il arrive aussi que l'étudiant doive choisir l'énoncé qui est le plus approprié du point de vue sociolinguistique. À l'item 11, deux réponses sont éliminées à cause d'un contresens mais le choix entre b et c est relié au registre:

You are in an elevator with many colleagues. Unfortunately, you spill some coffee on one of your colleagues' arms. What should you say?

a) Excuse-toi au moins.
b) Ayez l'obligeance de me pardonner.
c) Excuse-moi, je suis désolé.
d) Mes apologies.

Ce dernier item fait référence à un acte de parole particulier, s'excuser. Plusieurs items procèdent ainsi avec différents actes de parole. L'item 19 fait référence à l'expression d'une possibilité par rapport à celle d'un doute:

You are waiting for Maurice. Maurice is usually late. How could you say that you are almost sure he will be late?

a) Cela se peut qu'il soit en retard.
b) Il n'arrivera peut-être pas à l'heure.
c) Il est possible qu'il soit en retard.
d) Je doute beaucoup qu'il arrive à l'heure.

Il faut noter que tous les distracteurs sont grammaticalement corrects. Il ne s'agit donc pas pour l'étudiant de trouver l'énoncé correct mais plutôt d'identifier celui qui est sémantiquement et socialement acceptable. Il ne fait pas de doute que dans cette perspective, on risque de faire intervenir une multitude de considérations qui ne sont pas nécessairement pertinentes pour le classement des étudiants et qui peuvent compliquer la standardisation du test. On remarquera aussi que cette partie suppose la connaissance de l'anglais. Même si tous les sujets lisaient l'anglais et que nous avons essayé de formuler les situations le plus simplement possible, il ne fait pas de doute que l'habileté à lire en anglais est intervenue.

Le dernier sous-test exploite un type de tâche fort populaire dans les tests de langue seconde. Bien que tout à fait artificiel, l'exercice lacunaire permet la vérification d'une multitude d'éléments. On sait que dans une approche fondée sur l'exploitation d'une banque d'items, le test de closure pose des problèmes sérieux. Comme il nous semblait que la phrase à trou était susceptible de donner une information équivalente, nous avons souvent assimilé ce sous-test à un test de closure. La

sélection n'est pas aléatoire et on se conforme au format des choix multiples. Ici encore l'expérience du professeur de langue nous a guidé dans la sélection du mot à supprimer de même que la formulation des distracteurs. On peut mesurer des éléments de nature diverse. Ainsi l'item 27 vérifie la connaissance du vocabulaire de base:

Hier, c'était _____. Donc, aujourd'hui c'est jeudi.

a) jeudi b) vendredi
c) mercredi d) lundi

Par contre, l'item 14 vérifie nettement les connaissances grammaticales de l'étudiant:

– Avez-vous rencontré Pierre?
– Oui, nous _____ avons rencontré une fois.

a) en b) le
c) l' d) y

On peut également mesurer des aspects plus mécaniques qui sont souvent des indicateurs du degré de maîtrise. C'est le cas notamment de l'emploi des prépositions que cherche à mesurer l'item 22:

Le but du jeu est _____ lancer la balle dans le filet.
L'équipe qui marque le plus de buts gagne.

a) de b) à
c) dans d) pour

L'intérêt de ce type d'items tient aussi au fait que l'étudiant y répond rapidement, sans devoir fournir un grand effort de concentration. C'est pourquoi, cette partie termine le test.

Les 150 items qui composaient la version pré-expérimentale du test se trouvaient dans un cahier broché d'une vingtaine de pages. Chaque partie commençait avec un encadré où se trouvait formulée, succintement, la consigne. On demandait aux étudiants de ne pas écrire sur le questionnaire mais de se servir plutôt de la

feuille de réponses qui accompagnait le cahier. On avait joint à la feuille de réponses, une autre feuille que devaient lire et signer les étudiants: on leur expliquait les buts du test et l'usage qu'on ferait des résultats.

3.1.1.2 *L'échantillon*

Même si on entrevoyait des utilisations auprès de clientèles adultes ou de niveau secondaire, le test s'adresse plus particulièrement aux étudiants de niveau post-secondaire qui s'inscrivent dans des cours de français langue seconde. Plus spécifiquement, nous avions à l'esprit les besoins particuliers du programme de bourses du Secrétariat d'État. Comme nous l'avons déjà mentionné, ce programme offre la chance à des étudiants canadiens fréquentant des institutions secondaires et post-secondaires de s'inscrire à des sessions intensives de six semaines en langue seconde. Au cours de la session, les étudiants reçoivent au moins trois heures de cours de langue par jour, du lundi au vendredi. Le reste du temps est consacré à des activités organisées et des ateliers qui doivent se dérouler dans la langue seconde. Parce qu'on insiste sur l'usage de la langue seconde en tout temps, il s'établit entre la salle de classe et le milieu une dialectique favorable à l'acquisition de la langue. Cette situation pose des problèmes particuliers du point de vue de l'évaluation d'autant plus que s'y associent des considérations pratiques fort importantes. En effet, le défi est souvent de classer le plus adéquatement possible, en une demi-journée, un groupe de 40 à 500 étudiants de formations diverses et venant de différentes régions du pays.

L'Université York offre, hors campus, des sessions de français langue seconde, dans le cadre du programme du Secrétariat d'État. La session a lieu au printemps, à Saint-Georges de Beauce, une petite ville francophone située à une centaine de kilomètres au sud de Québec. Les étudiants habitent dans des familles, participent à diverses activités et suivent des cours crédités de l'Université York. Le programme de Saint-Georges offre plusieurs avantages du point de vue de l'expérimentation. Avec une centaine d'étudiants, il s'agit d'un programme de taille moyenne où l'on retrouve tous les niveaux,

des parfaits débutants jusqu'au plus avancés. Contrairement à beaucoup de programmes du même genre, la quasi-totalité des étudiants qu'on y admet sont des boursiers du Secrétariat d'État. On s'assure ainsi d'une bonne homogénéité de l'échantillon:

- l'âge des étudiants varie peu, la moyenne s'établissant à un peu plus de vingt ans;
- tous sont inscrits comme étudiants à plein temps dans une université canadienne;
- tous sont soit citoyens canadiens, soit immigrants reçus.

On peut donc compter sur une connaissance commune de l'anglais et sur une expérience du contexte culturel canadien. Cela est d'autant plus marqué pour le programme de Saint-Georges de Beauce que les deux tiers des participants viennent de la région métropolitaine de Toronto.

3.1.1.3 *Le déroulement de l'expérimentation*

Le programme de Saint-Georges permettait également de bien contrôler les conditions d'administration. Le test a été administré à 109 étudiants qu'on avait divisés en trois sous-groupes. À chaque sous-groupe était assigné un surveillant qui s'occupait de distribuer, de ramasser le matériel et de répondre aux questions des étudiants relativement à la consigne. Le surveillant était aussi chargé d'inscrire au tableau la correction de quelques erreurs mineures qui s'étaient glissées dans le questionnaire.

Conformément aux exigences du Secrétariat d'État, le test a été administré deux fois: le premier jour de la session (pré-test) et le dernier jour, six semaines plus tard (post-test). Comme nous cherchions d'abord à vérifier la valeur du test comme outil de classement, nous avons considéré uniquement les résultats au pré-test. Les résultats du post-test n'ont été utilisés que pour les quelques étudiants qui y avaient obtenu un score inférieur à celui de leur pré-test. Enfin, il faut souligner que la situation à Saint-Georges

se prêtait bien à l'expérimentation car pour les étudiants inscrits à York, le test n'avait pas de conséquence sur le classement réel. En effet, pour ces étudiants, le cours s'inscrivait dans une séquence de cours déjà prévue. Par contre, leurs scores servaient à établir des normes à partir desquelles nous avons pu classer les quelque 35 autres étudiants. Ce mode de classement a semblé d'ailleurs avoir été efficace car aucun des changements de groupe qu'on a effectués par la suite n'était attribuable à une erreur de classement.

Bien qu'on ait encouragé les étudiants à ne pas s'attarder à un item particulier, tous avaient assez de temps pour répondre. La contrainte de temps n'a donc joué aucun rôle mais il n'est pas exclu qu'un effet de fatigue soit intervenu. On avait alloué trois heures pour l'administration du test et la plupart des étudiants ont mis un peu plus de deux heures. Le premier sous-test s'est nettement avéré le plus long à faire.

3.1.2 *L'analyse*

3.1.2.1 *Les statistiques générales*

Les 150 réponses des 109 répondants ont été transcrites et rangées dans un fichier pour le traitement informatique. Ces données ont d'abord été soumises au programme CORREC, un programme que nous avons élaboré pour effectuer une première analyse des trois sous-tests. D'abord, le programme corrige les résultats des étudiants et les ordonne selon le type de tri requis; on obtient à la fois le score brut et le score standardisé (score *z*). CORREC fournit ensuite des statistiques générales sur le test: moyenne, variance, fiabilité (KR-20)... Enfin, on trouve une analyse sommaire des items tenant compte de la difficulté des items (la probabilité d'une bonne réponse) et de la distribution des réponses selon les distracteurs.

Nous reproduisons toutefois dans le tableau 3.1 les statistiques générales des trois sous-tests.

TABLEAU 3.1
Statistiques générales de la version 1

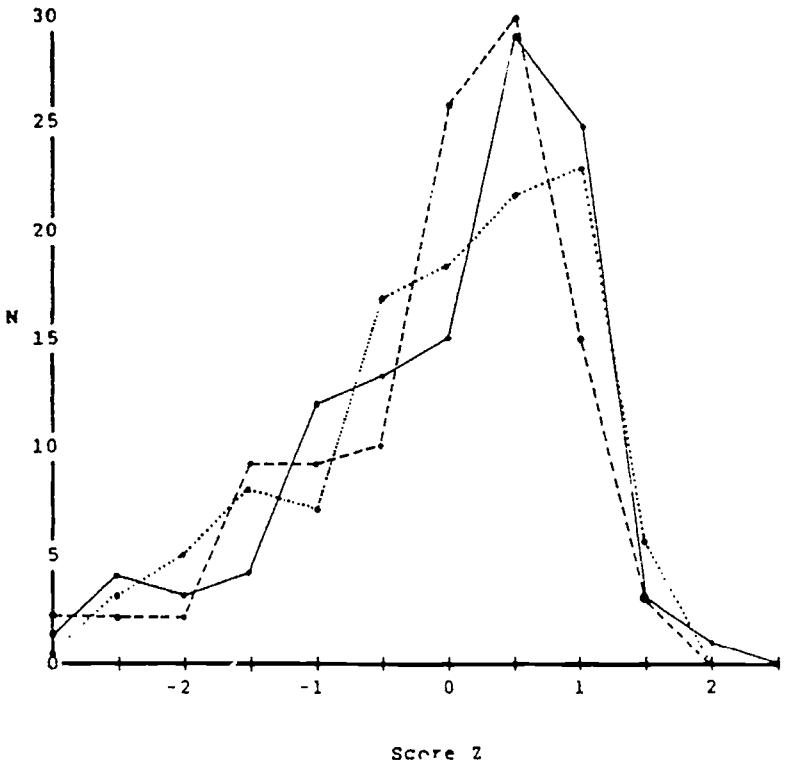
	#1	#2	#3	TOTAL
Maximum	46	44	46	133
Minimum	8	14	11	35
Moyenne	34.220	32.807	31.193	98.220
Variance	87.025	34.731	75.861	516.784
Écart-type	9.329	5.893	8.710	22.733
Fiabilité	0.903	0.756	0.886	0.956
Erreur-type	2.899	2.908	2.937	4.747

On observe tout d'abord que personne n'a obtenu de score parfait et ce malgré la présence d'étudiants avancés. On peut donc penser que certains items étaient beaucoup trop difficiles ou que les items plus difficiles discriminaient mal. Par ailleurs, ainsi qu'on pouvait s'y attendre avec un test à choix multiple, on n'observe aucun score nul: il est vraisemblable qu'en choisissant au hasard une des quatre réponses proposées, on obtiendra un score d'au moins 35 sur 150. D'après les moyennes, le sous-test de compréhension s'est avéré le plus facile et celui des phrases lacunaires, le plus difficile. Cet ordre est l'inverse de ce que nous attendions, mais il est possible qu'un effet de fatigue ait pu influencer les moyennes. En appliquant le test de Scheffé, on s'aperçoit que seule cette différence est significative ($p < .05$); les différences entre la moyenne du deuxième sous-test et celle des deux autres sous-tests peuvent être attribuées au hasard. Dans l'ensemble, la moyenne générale de 98.22 (65%) est tout à fait satisfaisante pour un test offrant quatre options pour chaque réponse. L'écart type (et conséquemment la variance) est plus grand au premier sous-test; il est toutefois beaucoup trop réduit au deuxième sous-test. On peut penser que la partie de compréhension d'un paragraphe discrimine mieux que celle où l'étudiant choisit l'énoncé approprié; dans ce dernier cas, les scores auraient tendance à se concentrer autour de la moyenne. Cette interprétation se trouve confirmé par l'indice de fiabilité KR-20 qui se calcule ainsi:

$$KR\text{-}20 = \frac{k}{(k-1)}\left[\frac{s_t^2 - \Sigma PQ}{s_t^2}\right] \qquad (3.1)$$

où k représente le nombre d'items, s_t^2 la variance des scores, P la probabilité d'une réponse correcte tandis que Q = P — 1. Le KR-20 est généralement une approximation satisfaisante de la fiabilité des tests de langue (Krzanowski et Woods 1984). Les indices de fiabilité se comparent avantageusement avec ceux calculés par Davidson (1988) dans le cas de tests de langue standardisés. L'erreur de mesure indique les bornes théoriques de l'intervalle de confiance: 2/3 des scores réels devraient se situer à l'intérieur de cet intervale. La marge d'erreur interdit donc un classement serré surtout pour avec le deuxième sous-test dont la variance est peu élevée.

FIGURE 3.1
Distribution des scores de la version 1



En portant les scores standardisés sur un graphique (figure 3.1), on est surpris de voir la forme leptokurtique de la courbe du premier sous-test alors que la variance de ces scores était la plus grande. Toutefois, ce qui est plus important encore c'est que les scores standardisés obtenus lors de cette pré-expérimentation, se concentrent autour du niveau «Intermédiaire fort» plutôt qu'autour de la moyenne. Même s'il est à peu près impossible d'obtenir une courbe symétrique dans un test où les sujets peuvent deviner la réponse, il n'est pas souhaitable dans un test de classement, que les scores s'agglutinent à une extrémité de l'échelle d'habileté. Dans les trois sous-tests, on pourrait rendre la distribution plus normale et améliorer la discrimination en ajoutant des items difficiles.

Comme le montre la figure 3.2, c'est le sous-test #1 qui bénéficierait davantage de l'addition d'items plus difficiles. L'histogramme obtenu en considérant les probabilités de réponse exacte pour chaque item montre en effet que la plupart se situent entre .9 et .6.

FIGURE 3.2
Répartition des items du sous-test #1 par degré de difficulté (version 1)

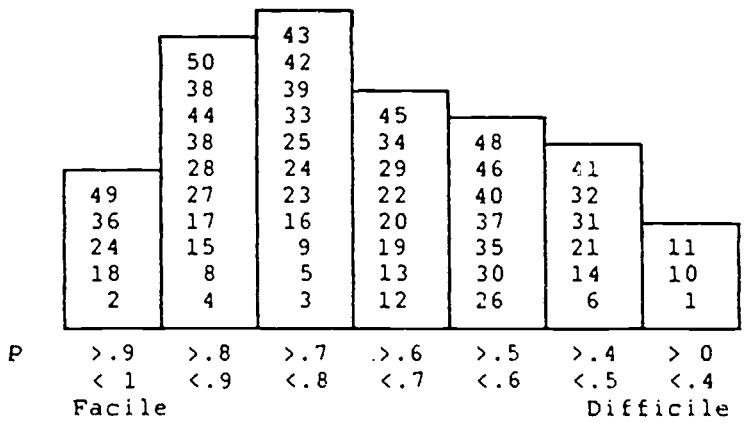


FIGURE 3.3
Répartition des items du sous-test #2 par degré de difficulté (version 1)

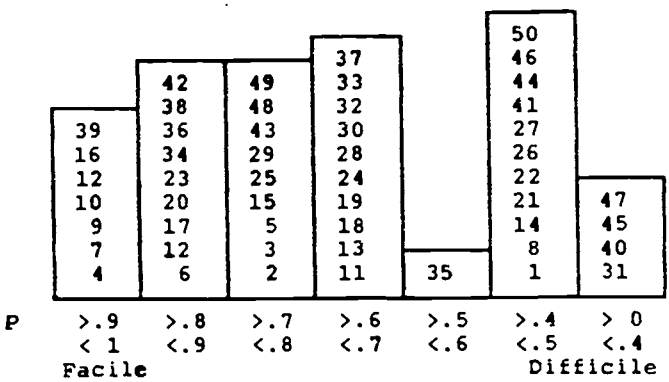
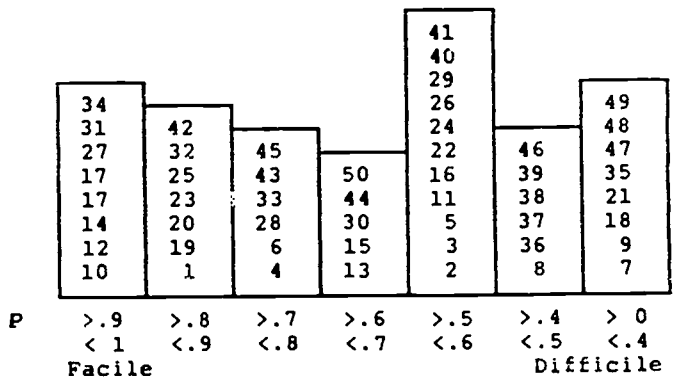


FIGURE 3.4
Répartition des items du sous-test #3 par degré de difficulté (version 1)



Les niveaux de difficulté sont beaucoup mieux distribués pour le sous-test #3 (figure 3.4) ce qui explique la forme plus normale de la courbe des scores standardisés. Quant au sous-test #2 (figure 3.3), il semble qu'il y ait une démarcation entre les items faciles d'une part et les items difficiles d'autre part, et que le faible indice de fiabilité cache un problème au plan de la discrimination des items.

3.1.2.2 *Les corrélations*

À l'aide d'un programme que nous avons créé, le programme COREVAR, nous avons produit les matrices de corrélation, de corrélation au carré et de variance/covariance. L'examen des covariances du tableau 3.2a nous amène à suspecter des problèmes au plan de la discrimination des items du sous-test #2 puisque la variance de ce sous-test est le nombre le plus petit de la matrice. Les scores tendraient donc à se concentrer autour de la moyenne.

TABLEAU 3.2
Corrélations et covariances de la version 1

a) Covariances entre les scores

	# 1	# 2	# 3
#1 Compréhension	87.025	45.395	71.337
#2 Énoncé approprié	45.395	34.731	42.852
#3 Phrases à trou	71.337	42.852	75.861

b) Corrélations: coefficient r de Pearson

	# 1	# 2	# 3
#1 Compréhension	1.000	0.826	0.878
#2 Énoncé approprié	0.826	1.000	0.835
#3 Phrases à trou	0.878	0.835	1.000

c) Carrés des coefficients de corrélation

	# 1	# 2	# 3
#1 Compréhension	1.000	0.682	0.771
#2 Énoncé approprié	0.682	1.000	0.697
#3 Phrases à trou	0.771	0.697	1.000

Cependant, ce qui frappe davantage, ce sont les fortes corrélations que l'on observe entre les sous-tests dans le tableau 3.2b. Tous les coefficients de corrélation sont supérieurs à .8. La corrélation entre le test de compréhension et celui des phrases à trou est de .878 ce qui veut dire, comme l'indique le tableau 3.2c, que ces

deux tests ont en commun 77% de leur variance. Plus étonnant encore, les corrélations impliquant le sous-test #2, sont supérieures à l'indice de fiabilité de ce sous-test! Si la division en trois sous-tests se justifiait de par la nature de la tâche, il n'en reste pas moins que ces parties du test semblent mesurer principalement des aspects communs de la maîtrise de la langue. Gardant à l'esprit le fait que la décision reliée au classement est essentiellement unidimensionnelle, il nous apparaissait tout à fait justifié de conserver ces trois sous-tests dans les versions ultérieures.

En utilisant le sous-programme *SCATTERGRAM* de *SPSS-X*[1] (Nie *et al.* 1983), nous avons pu obtenir les diagrammes de dispersion. Ces diagrammes montrent que la corrélation entre chaque sous-test et le test total est linéaire. On note une concentration des points dans la partie supérieure des diagrammes ce qui confirme qu'il faudrait ajouter des items plus difficiles. Les points s'écartent généralement peu de la droite puisque les coefficients de corrélation entre les parties et l'ensemble sont assez élevés: .961 pour le sous-test #1, .918 pour le sous-test #2 et .960 pour le sous-test #3.

3.1.2.3 *L'analyse des items*

Le relevé de CORREC donne peu d'information sur la discrimination de chaque item. Afin de déterminer quels étaient les items les plus discriminants c'est-à-dire ceux qui départageaient le mieux les sujets, nous avons utilisé le programme *LERTAP*[2]. En plus de la distribution des réponses et de l'indice de probabilité qui y est associé, *LERTAP* fournit des indices sur le comportement de l'item par rapport à l'ensemble du sous-test et du test complet. Tant pour la bonne réponse que pour les distracteurs, on trouve deux coefficients: la corrélation point-bisérielle et la corrélation bisérielle. Ces deux coefficients fournissent la même

[1] *SPSS-X* est installé sur l'ordinateur VAX de l'Institut d'études pédagogiques de l'Ontario.

[2] Le programme *LERTAP* est installé sur l'ordinateur VAX de l'Institut d'études pédagogiques de l'Ontario.

information c'est-à-dire la corrélation entre une réponse dichotomique (exact/inexact) et un score. Toutefois, le calcul de la corrélation bisérielle, en supposant une distribution normale pour des variables dichotomiques, produit des coefficients plus élevés surtout lorsque les indices de probabilité prennent des valeurs extrêmes. *LERTAP* fournit également la moyenne (au sous-test et au test complet) des sujets qui ont choisi la bonne réponse et chaque distracteur.

Nous avons fait remarquer que les items du sous-test de compréhension étaient souvent trop faciles mais que dans l'ensemble ils semblaient bien discriminer. L'item 25 de ce premier sous-test consistait en un court message sur la prévention des maux de dos; il s'avérait particulièrement discriminant bien qu'un peu facile. Pour cet item, le relevé de *LERTAP* fournissait l'information suivante[3]:

```
TEST NO 1    VAL1.DAT              SUBTEST 1      COMPREHENSION
ITEM NUMBER 25      COEFFICIENTS OF CORRELATION      MEANS

OPTION WT   N     P     PB-ST PB-TT   B-ST  B-TT     ST      TT
    1  0   11   10.1    -0.24 -0.23  -0.41 -0.39   27.64   82.73
    2  0    6    5.5    -0.33 -0.35  -0.69 -0.72   21.33   65.50
    3  0   13   11.9    -0.41 -0.39  -0.66 -0.63   24.00   74.23
  C 4  1   77   70.6 C   0.65  0.65   0.86  0.86 C 38.12  107.75
    9  0    2    1.8    -0.13 -0.17  -0.38 -0.50   25.50   70.50
TOTAL     109
```

Par ailleurs, on a constaté des lacunes dans le deuxième sous-test, au plan de la discrimination. Ainsi à l'item 26, l'étudiant doit choisir l'énoncé qui convient pour avertir quelqu'un de faire attention à une marche. Peut-être à cause de la difficulté lexicale que représente le mot «marche» dans l'énoncé correct «Attention à la marche», les étudiants les plus avancés choisissent plutôt le faux-ami structural «Surveille tes pas». La corrélation entre la réponse correcte et le score global est donc à peu près nulle.

---

[3] Pour des raisons d'espace, nous ne reproduisons pas le relevé mécanographique des analyse de *LERTAP* et nous nous limitons à un exemple typique pour chaque sous-test. Toutefois, ces documents sont disponibles sur demande.

TEST NO 1 VAL1.DAT SUBTEST 2 ENONCE APPROPRIE
ITEM NUMBER 26 COEFFICIENTS OF CORRELATION MEANS

OPTION	WT	N	P		PB-ST	PB-TT	B-ST	B-TT		ST	TT
1	0	16	14.7		0.28	0.30	0.43	0.46		36.75	114.50
2	0	19	17.4		-0.19	-0.16	-0.28	-0.24		30.42	90.11
3	0	27	24.8		-0.11	-0.13	-0.15	-0.18		31.70	93.07
C 4	1	47	43.1	C	0.04	0.03	0.05	0.06	C	33.06	98.91
9	0	0	0.0		0.00	0.00	0.00	0.00		0.00	0.00
TOTAL		109									

Dans certains cas, il apparaissait que la situation avait été interprétée différemment; pour d'autres items il semblait que les sujets ignoraient comment réagir dans la situation présentée. Bref, c'est nettement dans cette partie que nous avons observé les coefficients de discrimination les plus bas.

Les réponses au sous-test #3 (phrases à trou) étaient beaucoup plus prévisibles, ce qui a simplifié l'analyse.

TEST NO 1 VAL1.DAT SUBTEST 3 PHRASES A TROU
ITEM NUMBER 2 COEFFICIENTS OF CORRELATION MEANS

OPTION	WT	N	P		PB-ST	PB-TT	B-ST	B-TT		ST	TT
1	0	3	2.8		-0.12	-0.04	-0.31	-0.11		25.00	92.33
2	0	37	33.9		-0.25	-0.19	-0.32	-0.24		28.22	92.24
3	0	5	4.6		-0.38	-0.42	-0.82	-0.92		16.20	54.60
C 4	1	63	57.8	C	0.46	0.40	0.58	0.51	C	34.62	105.97
9	0	1	0.9		-0.14	-0.13	-0.52	-0.51		19.00	67.00
TOTAL		109									

L'item 2 où l'étudiant doit choisir la préposition qui accompagne le verbe «finir» («fini ***de*** manger») illustre ce que nous avons observé dans la plupart des items de cette partie: une question de difficulté moyenne, assez discriminante.

En complément aux résultats du programme *LERTAP*, nous avons traité les données avec le programme *BICAL*[4], un programme de calibration des items selon le modèle de Rasch. L'utilisation de ce programme à ce stade se justifiait à deux points de vue. D'une part,

[4] Le programme *BICAL* est installé sur l'ordinateur VAX de l'Institut d'études pédagogiques de l'Ontario.

nous n'avions pas assez de sujets pour utiliser le modèle à trois paramètres que nous comptions utiliser par la suite. D'autre part, *BICAL* fournit des renseignements qui peuvent souvent faciliter l'analyse des items d'un test de langue (Perkins et Miller 1984). Ainsi, le programme divise l'échantillon en sous-groupes selon l'habileté et inscrit les proportions de réponses correctes observées. De plus, le programme calcule non seulement la corrélation point-bisérielle mais aussi un indice de discrimination obtenu à partir des données de la calibration. Enfin, le programme donne la difficulté de l'item (en logits) de même qu'un indice d'adéquation qui montre dans quelle mesure les réponses coïncident avec la courbe caractéristiques de l'item. Par ailleurs, *BICAL* peut éliminer, avant le calcul final, un certain nombre de sujets qui ont des configurations de réponses marginales (*misfitting response patterns*).

Comme le test avait été aussi administré au terme des six semaines du programme, nous nous sommes interrogés sur le fait que onze étudiants avaient obtenu, au post-test, un score inférieur à celui du pré-test. On aurait pu y voir l'indication d'items peu fiables. Toutefois, nous avons été très prudents dans l'utilisation de ces données pour l'analyse des items car les conditions d'administration du post-test était loin d'être idéales de sorte que les écarts semblaient plutôt attribuables à l'inattention et à la fatigue. En effet, l'écart par item chez ces onze étudiants était plus manifeste vers la fin du test ou avec des items qui demandaient plus de concentration de leur part. Le fait que les onze sujets étaient plutôt avancés (moyenne de 116, au pré-test) indiquait bien que les items échoués à la deuxième administration étaient les plus difficiles ou les plus exigeants sur le plan cognitif.

Nous avons finalement complété l'analyse des items en administrant le test à trois étudiants francophones qui ne se spécialisaient pas en français ou dans l'enseignement. En étudiant les résultats de locuteurs natifs au *TOEFL*, Angoff et Sharon (1971) ont trouvé qu'un test conçu pour l'évaluation d'une langue étrangère pouvait difficilement servir à faire des distinctions entre des locuteurs natifs. Par contre, Friedman (1984) signale que ces données peuvent être fort utiles pour valider un test, notamment

pour la détection de biais culturels. Il faut noter que nos trois étudiants francophones étaient inscrits dans une université anglophone, de sorte qu'on pouvait penser *a priori* que les différences culturelles ou la maîtrise de l'anglais intervenaient peu. Les scores obtenus étaient de 142, 133 et 140. Les items où au moins deux des trois francophones avaient failli ont été revus et la plupart ont été corrigés ou remplacés.

3.1.3 ***Sommaire des modifications***

3.1.3.1 *Compréhension*

Comme le premier sous-test semblait un peu trop facile, certains items dont la probabilité de réponse correcte était très élevée ont été remplacés par des items plus difficiles. C'est le cas des items 2, 24 et 49. Le texte de l'item 36 a été modifié pour que la réponse soit moins évidente.

Même si la discrimination était en général assez bonne, les items 11, 13, 21 et 37 ont dû être remplacés parce qu'ils étaient peu efficaces pour départager les étudiants. On les a remplacés par des items plutôt difficiles. Les distracteurs de l'item 1 ont tous été reformulés afin d'améliorer la discrimination.

Pour d'autres items (10, 15 et 32), les corrections se sont limitées à un ou deux distracteurs qui semblaient peu efficaces. Enfin, signalons que l'item 9 a été modifié parce qu'il ressemblait trop à l'item 42. Au total, 13 items (26%) ont été révisés; parmi eux 7 (14%) ont été remplacés.

3.1.3.2 *Énoncé approprié*

Les changements au deuxième sous-test ont été beaucoup plus nombreux. Même si le niveau de difficulté était juste, l'item 7 a été révisé parce qu'il était trop facile alors que les items 40 et 50 ont été remplacés parce que trop peu d'étudiants y répondaient correctement.

Plusieurs items étaient nettement déficients du point de vue de leur discrimination. On a carrément remplacé les items 17, 22, 26, 28 et 42. Ce dernier était d'ailleurs tout à fait inadéquat par rapport au modèle à un paramètre. D'autres ont été modifiés, parfois substantiellement (items 27, 29, 31, 35 et 37).

Dans l'espoir de voir s'améliorer la discrimination on a également révisé un ou deux distracteurs dans les items suivants: 4, 11, 13, 14, 15, 16, 23, 24, 29, 39, 47 et 49.

Afin d'éliminer des ambiguïtés dans l'interprétation de la situation, on a apporté des changements mineurs dans la formulation de quelques situations: c'est le cas des items 18 et 33.

En comptant le nombre d'items qui ont subi des transformations (26 items soit 52%), on voit que moins de la moitié sont restés intacts. Cependant, seulement 7 (14%) items ont été remplacés. On peut donc voir que pour ce type de questions où l'on fait appel autant à des jugements sémantiques ou à une norme sociale qu'à une norme linguistique stricte, le processus d'analyse des items est fort important.

3.1.3.3 *Phrases à trou*

Dans l'ensemble, les changements au troisième sous-test étaient plutôt des modications que des substitutions. Ainsi on ajusté le niveau de difficulté en modifiant les items 17 et 31 (trop faciles) et les items 18 et 35 (trop difficiles). Par contre, l'item 10 était beaucoup trop facile et a été remplacé. Sept items semblaient moins efficaces du point de vue de la discrimination. On a pu modifier les items 16, 27, 36, 37 et 45 mais il a fallu remplacer les items 24 et 39. Notons d'ailleurs que les items 36, 37 et 39 s'intégraient mal au modèle de *BICAL*. Au cours de ces changements, on a généralement élargi les contextes afin de ne permettre qu'une seule réponse.

Seulement trois items ont vu un ou plusieurs distracteurs modifiés: 3, 43 et 50.

Dans l'ensemble les révisions ont affecté 15 items (30%). De ce nombre, seulement 3 (6%) ont été remplacés.

3.2 De la version expérimentale aux versions finales

3.2.1 *L'expérimentation*

3.2.1.1 *La cueillette des données*

À la suite des changements effectués au test, celui-ci a été reproduit en vue d'une expérimentation à grande échelle. On trouvera quatre questions où on demandait à l'étudiant d'évaluer lui-même son niveau de français et de préciser comment il avait atteint ce niveau. Il suffisait à l'étudiant de cocher les cases appropriées ou d'indiquer le nombre d'années d'études du français.

La feuille de réponse était accompagnée d'une autre feuille expliquant le but du test; 991 étudiants ont signé cette feuille signifiant ainsi qu'ils consentaient à ce qu'on utilise les résultats à des fins d'expérimentation. Comme l'expérimentation s'est déroulée à une beaucoup plus grande échelle que la pré-expérimentation, il nous a fallu visiter plus d'un établissement. Toutefois, nous nous en sommes tenu à des cours intensifs qui s'alignaient sur les objectifs généraux du programme de bourses du Secrétariat d'État. Une dizaine d'établissements ont accepté de participer à l'expérimentation. Certains ont utilisé les résultats du test pour leur propre classement. La grande majorité des étudiants étaient des boursiers du programme du Secrétariat d'État. Quelques-uns participaient à des sessions intensives sans être boursiers; d'autres s'étaient inscrits dans un cours régulier dispensé par un établissement post-secondaire. Cependant, dans tous les cas, nous avons veillé à ce que les caractéristiques générales des étudiants boursiers se retrouvent chez tous les sujets. Nous avons donc dû, dès le départ, retirer quelques étudiants qui ne répondaient pas à cette exigence: c'est le cas notamment des étudiants étrangers

dont l'anglais n'était pas la langue maternelle ou de certains sujets plus âgés. Le millier d'étudiants qui ont accepté de faire le test se répartissait ainsi:

Collège de Saint-Boniface (boursiers)	65
Collège de Saint-Boniface (programme MIELS)	81
Collège de Rivière-du-Loup	36
Université de Moncton	39
Collège Georges Brown à La Pocatière	92
Université Carleton (sessions d'été)	88
Université Carleton (inscriptions de septembre)	49
Université York à Saint-Georges (pré-test)	97
Université York à Saint-Georges (post-test)	92
Collège Bois-de-Boulogne	122
Université Laurentienne à Sudbury	68
Centre linguistique du Collège de Jonquière	101
Université Western à Trois-Pistoles	61

Quelques établissements ont administré eux-mêmes les tests et nous ont fait parvenir les feuilles de réponses. Toutefois, dans la plupart des cas, nous nous sommes rendus sur place pour administrer le test. Il faut souligner que les directeurs, les coordonnateurs pédagogiques, les professeurs et les étudiants se sont toujours bien prêtés à l'expérimentation. Le nombre de sujets et la qualité des conditions d'administration sont largement attribuables à cette heureuse collaboration. Plus de 400 sujets ont répondu au test complet: on les retrouve surtout à l'Université York, au Collège de Saint-Boniface, à l'Université Laurentienne et à l'Université Carleton. Par ailleurs, il n'était pas toujours possible de consacrer deux heures et demie à un test, surtout si les résultats ne servaient pas au classement. C'est pourquoi, on a souvent dû limiter l'expérimentation à une ou deux parties en s'assurant que les résultats partiels comptent un nombre à peu près égal de sujets pour chaque sous-test.

Nous avons réuni ces données pour créer deux échantillons: l'échantillon d'analyse et l'échantillon de calibration. L'échantillon d'analyse devait servir aux statistiques générales, à l'analyse classique des items et aux corrélations. Il était donc important que la représention des niveaux reflète celle de la population visée par le

test. On n'a conservé que les établissements où les trois parties du test avaient été administrées à tous les étudiants inscrits. De plus, on a éliminé les post-tests de l'Université York puisque le fait que les mêmes sujets aient repris le test après six semaines intensives risquait de fausser la représentativité de l'échantillon. Il restait donc dans l'échantillon d'analyse, 328 sujets. Ces mêmes sujets ont été intégrés au deuxième échantillon, l'échantillon de calibration, qui devait servir à la paramétrisation des items. En vertu du principe d'invariance, le processus de calibration ne requiert pas une distribution normale de l'habilité mais exige en contrepartie un grand nombre de sujets, surtout avec un modèle à trois paramètres. À condition que tous les niveaux soient représentés, une légère surreprésentation ou sous-représentation d'un niveau ne risque pas de fausser les paramètres. Nous avons d'abord ajouté les données du post-test de l'Université York. Nous avons ensuite composé des tests complets à l'aide des résultats partiels c'est-à-dire des réponses des étudiants à qui on n'avait administré qu'un ou deux sous-tests. Pour ce faire, nous avons utilisé le programme COREVAR pour calculer les covariances et établir des tables de régression pour chaque sous-test par rapport aux deux autres sous-tests. Par exemple, un score de 22 au premier sous-test devait être joint à un score de 28 au deuxième sous-test et de 22 au troisième; un score de 35 au troisième sous-test correspondait, pour les premier et deuxième à 39 et 36 respectivement. Ces tables ont servi de guide pour créer un ensemble de tests composites dont les moyennes s'approchaient de celle de l'échantillon d'analyse. L'opération a permis de constituer un échantillon de calibration comprenant l'équivalent de 749 sujets auxquels avaient été administrées les 150 questions du test complet. Contrairement à l'échantillon d'analyse, l'échantillon de calibration de chaque sous-test n'obéissait pas nécessairement à une distribution normale mais se comparait à l'échantillon d'analyse à tout autre point de vue.

3.2.1.2 *L'épuration des données*

Une fois la laborieuse étape de la saisie des données achevée, nous avons procédé à quelques analyses préliminaires afin de

détecter les sujets qui risquaient davantage de contaminer les données plutôt que de les compléter. De fait, nous souhaitions ne pas devoir retirer un trop grand nombre de sujets car 750 représente un échantillon restreint pour une calibration utilisant trois paramètres. Toutefois, il nous semblait également qu'en retirant les configurations de réponses divergentes, la perte de quelques sujets serait largement compensée par un gain appréciable quant à la fiabilité des données.

Nous avons tenu compte de quatre types de critères en ce qui a trait au retrait de certains sujets.

- Le nombre de réponses: La plupart des logiciels distinguent entre les réponses omises (ou annulées) et l'absence de réponse à cause d'un manque de temps ou d'un abandon. Nous estimions cependant que lorsque qu'il manque plus de la moitié des réponses le sujet risque d'apporter des données peu fiables et nous avons éliminé ces cas. De même, les sujets qui donnaient plus de 25 réponses identiques ou une série ininterrompue d'une dizaine de réponses identiques ont été retirés du fichier.

- Le score: Comme l'algorithme de calibration des items s'accommode mal de scores parfaits ou nuls, ces cas ont été retirés. De fait, en prenant en considération l'effet du hasard dans un test comprenant des items à quatre choix, nous n'avons pas inclus dans les données tout score inférieur à 4.

- *BICAL:* Nous avons demandé au programme *BICAL* d'identifier les sujets qui montraient un indice d'inadéquation élevé par rapport au modèle de Rasch ($t < 2.0$). La détection de ces cas est plus difficile et parfois impossible avec les logiciels qui font des analyses à trois paramètres. De la sorte, on a retiré 16 sujets du premier sous-test, 25 du deuxième et 37 du troisième.

- Les échelles implicationnelles: Cliff (1983) a proposé une série d'indices permettant d'évaluer dans quelle mesure la distribution des réponses se conforme au modèle implicationel de Guttman. Il conclut que, sans rivaliser avec la puissance et l'élégance des solutions obtenues avec la théorie du trait latent, cette approche est utile et évite de devoir se fonder sur les postulats l'analyse classique. Cliff *et al.* (1978) ont même proposé d'utiliser ces échelles pour l'administration d'un test adaptatif. Nous avons mis au point le programme SCALE.afin de construire des échelles implicationnelles pour chaque sous-test. Le programme ordonne, d'une part, les items selon leur difficulté et, d'autre part, les sujets selon leur score. Les indices de la colonne de gauche indiquent la proportion de réponses de chaque sujet qui obéissent au modèle implicationnel en tenant compte de l'effet du hasard; la dernière ligne du tableau fournit le même type d'indice au plan des items. Si le score était inférieur à 13, c'est-à-dire en deçà de ce qu'on obtient en répondant de façon purement aléatoire, nous avons éliminé tous les sujets qui présentaient des indices inférieurs .83. Si, comme c'était généralement le cas, un sujet obtenait 13 ou plus, nous ne l'avons éliminé que si l'indice ne dépassait pas .65.

Après l'épuration des données, il restait 314 sujets dans l'échantillon d'analyse. En ce qui concerne l'échantillon de calibration, on a conservé 695 sujets au premier sous-test, 683 au second et 661 au dernier.

3.2.2 *L'analyse*

3.2.2.1 *Les statistiques générales*

Les corrections apportées au test original ont permis d'améliorer la qualité de la mesure. Les moyennes, qui étaient un peu trop élevées, particulièrement pour le premier sous-test, ont diminué. Les variances qu'on espéraient voir augmenter surtout au

deuxième sous-test ont effectivement augmenté. Pour avoir une idée juste des hausses, il faut comparer l'administration de la version pré-expérimentale à Saint-Georges en 1986 avec l'administration de la version expérimentale dans le même établissement l'année suivante puisque le niveau des étudiants de Saint-Georges est légèrement plus avancé que celui de la majorité des autres établissements qui participent au programme de bourses. Par ailleurs, nous savons que les caractéristiques du groupe de Saint-Georges n'ont guère changé d'une année à l'autre. Le tableau 3.3 montre comment se comparent les moyennes et les écarts types des deux administrations.

TABLEAU 3.3
Moyennes et écarts types des versions 1 et 2

	1986: Version 1.1		1987: Version 2.2	
	Moyenne	Écart-type	Moyenne	Écart-type
#1 Compréhension	34.22	9.33	31.92	11.91
#2 Énoncé approprié	32.88	5.89	31.19	8.84
#3 Phrases à trou	31.19	8.71	28.23	11.21
Test complet	98.22	22.73	91.33	30.87

Nous reproduisons les statistiques générales dans le tableau 3.4.

TABLEAU 3.4
Statistiques générales de la version 2

	#1	#2	#3	TOTAL
Maximum	50	46	46	139
Minimum	0	0	0	5
Moyenne	29.591	30.372	25.756	85.720
Variance	141.123	66.546	104.815	827.958
Écart-type	11.88	8.158	10.238	28.774
Fiabilité	0.928	0.850	0.904	0.966
Erreur-type	3.18	3.16	3.166	5.343

On note que la moyenne du sous-test #3 est nettement inférieure à celle des autres sous-tests. Celà tient à la fois à la difficulté même du test et au fait que certains sujets n'ont pas terminé l'épreuve. On voit aussi que le sous-test #2 reste encore moins fiable que les deux autres bien que les changements l'aient beaucoup amélioré. Le coefficient de fiabilité (KR-20) de la première et de la dernière partie dépasse .9, ce qui est plus que satisfaisant avec 50 items. La fiabilité générale se situe à .97. Ce chiffre correspond à ce que Davidson (1988) a calculé avec un échantillon de 5 000 sujets au *TOEFL* de 1985 (146 items). Cela dépasse même la fiabilité d'un instrument comme le *CanTEST* (153 items) qui a été utilisé avec des étudiants chinois en séjour au Canada et dont la marge d'erreur s'est avéré tout à fait acceptable (DesBrisay, communication personnelle). Enfin, il convient de souligner le fait que l'erreur type est identique d'un sous-test à l'autre, soit 3.2.

TABLEAU 3.5
Moyennes et écarts types des deux échantillons

	Échantillon d'analyse		Échantillon de calibration	
	Moyenne	Écart-type	Moyenne	Écart-type
#1 Compréhension	29.60	11.88	30.96	11.60
#2 Énoncé approprié	30.37	8.16	31.19	8.23
#3 Phrases à trou	25.75	10.23	26.26	10.37
Test complet	88.1	29.01	85.72	28.77

Comme le montre le tableau 3.5, les moyennes sont légèrement supérieures (peut-être à la suite de l'inclusion des post-tests de York) et les variances sont comparables. Quant aux coefficients de fiabilité des deux échantillons, ils sont égaux.

3.2.2.2 *Les corrélations*

3.2.2.2.1 Les corrélations entre les sous-tests

L'amélioration de la fiabilité que nous avons constatée implique une réduction de la marge d'erreur. Or comme les erreurs

de mesure de différents tests ont la propriété de ne pas être corrélées, on peut penser que les corrélations entre sous-tests augmenteront. De fait, en examinant les corrélations du tableau 3.6, on voit que celles-ci ont à peine augmenté. Certes, la corrélation entre les première et deuxième parties a augmenté mais celle entre les deuxième et troisième parties a diminué un peu. Il faut examiner la matrice des variances et covariances pour constater l'amélioration entre les deux versions du test. Comme auparavant, le sous-test #1 montre le maximum de variance et le sous-test #2 le minimum de variance; cependant, les covariances ont augmenté partout.

TABLEAU 3.6
Corrélations et covariances entre les sous-tests de la version 2

a) Covariances entre les scores

	# 1	# 2	# 3
#1 Compréhension	130.379	76.106	98.459
#2 Énoncé approprié	76.106	60.389	62.098
#3 Phrases à trous	98.459	62.098	97.128

b) Corrélations: coefficient r de Pearson

	# 1	# 2	# 3
#1 Compréhension	1.000	0.858	0.875
#2 Énoncé approprié	0.858	1.000	0.811
#3 Phrases à trou	0.875	0.811	1.000

c) Carrés des coefficients de corrélation

	# 1	# 2	# 3
#1 Compréhension	1.000	0.736	0.766
#2 Énoncé approprié	0.736	1.000	0.657
#3 Phrases à trou	0.766	0.657	1.000

Par ailleurs, il est difficile de trouver des corrélations plus élevées compte tenu des coefficients de fiabilité qui ont été calculés. En effet, en appliquant la formule de correction pour

l'atténuation, formule créée dans le cadre de la théorie classique, on trouve des corrélations entre les scores véritables qui s'approchent et même dépassent la limite théorique de 1. L'équation a la forme suivante:

$$r_{T_x T_y} = \frac{r_{xy}}{\sqrt{r_{xx'} r_{yy'}}} \qquad (3.1)$$

Entre le premier sous-test et le second, on obtient un coefficient de .97, entre le premier et le troisième, un coefficient de 1.04 et enfin, entre le second et le troisième, un coefficient de 1.06. Ces résultats pour le moins surprenants indiquent probablement une estimation trop conservatrice des indices de fiabilité mais suggèrent aussi que les trois sous-tests évaluent un facteur commun.

3.2.2.2.2 L'analyse de *LISREL*

Afin d'explorer la structure factorielle du test nous avons cherché à vérifier la cohérence interne des sous-tests et à déterminer si l'hypothèse de tests congénériques pouvait être retenue (Jöreskog 1971, Linn et Werts 1979). Pour définir le test congénérique, nous définissons les variables suivantes:

X_{aj} : Le score observé pour le sujet a au test j
M_j : La moyenne du test j
b_j : Le coefficient de régression de X_j sur T
T_a : Le score véritable normalisé
E_a : L'erreur normalisée
n : Le nombre de sujets

Nous posons comme équation de base au test congénérique, l'équation 3.2:

$$X_{aj} = M_j + b_j T_a + E_{aj} \qquad (3.2)$$

Dans le cas où les deux tests j et j', mesurent la même variable latente T, on a aussi l'équation 3.3:

$$X_{aj'} = M_{j'} + b_{j'} T_a + E_{aj'} \qquad (3.3)$$

Par ailleurs, on obtient la sommation des produits ainsi:

$$\sum_{a=1}^{n} X_{aj} X_{aj'} = \sum_{a} (M_j + b_j T_a + E_{aj})(M_{j'} + b_{j'} T_a + E_{aj'}) \quad (3.4)$$

S'il s'agit de test parallèles, j = j' et 3.4 se réduit à l'équation 3.5:

$$\sum_{a=1}^{n} X_{aj} X_{aj'} = nM_j^2 + nb_j^2\sigma_T^2 + n\sigma_{E_j}^2 \quad (3.5)$$

Le programme *LISREL*[6] (Jöreskog et Sörbom 1983) est essentiellement conçu en vue d'analyses confirmatoires bien que l'ayons utilisé dans une perspective exploratoire. Nous voulions en effet vérifier le modèle le plus simple (un facteur unique) et ajouter progressivement des composantes jusqu'à ce que le modèle soit satisfaisant. *LISREL* nous permettait de poursuivre l'analyse jusqu'au test congénérique le plus complexe. De fait, *LISREL* est un programme qui sert à l'estimation des coefficients dans des équations structurales linéaires dont les applications sont très variées (Everitt 1984). Ainsi, Nelson *et al.* (1984) ont utilisé *LISREL* afin de valider un modèle d'acquisition à partir de résultats de tests. Le programme utilise des variables observables (des résultats de test, par exemple) et des variables latentes (les scores véritables, par exemple). Il permet de mesurer ces variables de même que l'erreur de la mesure qui leur est associée. Ce modèle postule qu'il y a un lien causal entre les variables observables et les variables latentes qui leur sont sous-jacentes.

On utilise les symboles suivants pour désigner un ensemble de vecteurs et de matrices:

ε et δ : vecteurs d'erreurs de mesure (non correllé avec η ni avec ξ);

ζ : vecteur de valeurs résiduelles (non correllé avec ξ);

[6] Le programme *LISREL* est installé sur l'ordinateur VAX de l'Institut d'études pédagogiques de l'Ontario.

$\hat{y}$ et $\hat{x}$: matrices de coefficients de régression de y sur η et de x sur ξ, respectivement;

η et ξ : matrices de variables latentes;

B et Γ : matrices de coefficients reliant η à η' et η à ξ, respectivement.

Ces symboles servent à définir deux types de modèles

- le modèle de l'équation structurale: $\eta' = B\eta + \Gamma\xi + \zeta$ (3.6)

- les modèles de mesure, pour y: $y = \Lambda y + \varepsilon$ (3.7a)
  pour x: $x = \Lambda x + \delta$ (3.7b)

LISREL établit une matrice de covariance Σ dont les éléments sont fonctions des matrices de coefficients ($\hat{y}$, $\hat{x}$, **B** et Γ), des matrices de variables de l'équation structurale (η et ξ) et des matrices des erreurs de mesure (ε et δ).

Nous avons d'abord divisé chaque sous-test en deux parties de 25 items chacune dans l'intention de calculer les corrélations entre les scores obtenus à chaque moitié de chaque sous-test. D'une part, le fait de disposer de six variables plutôt que de trois augmentait le nombre de degrés de liberté. D'autre part, on pouvait ainsi vérifier la cohérence interne de chaque sous-test. Nous aurions pu diviser les sous-tests entre items faciles et items difficiles mais un facteur de difficulté aurait obscurci les résultats. Nous aurions pu aussi comparer les 25 premiers items aux 25 derniers mais un facteur de lassitude aurait pu intervenir. Nous avons finalement décider de séparer les items désignés par un nombre impair et de ceux désignés par un nombre pair. On a donc réuni les items 1, 3 ... 49 de chaque sous-test puis les items 2, 4 ... 50. Le tableau 7 montre les corrélations entre les scores que nous avons obtenues avec COREVAR.

En examinant la matrice, on voit que les moitiés de chaque sous-tests sont fortement corrélées ($r > .99$). Ces valeurs donnent une idée d'une autre forme de fiabilité (*Split-half reliability*, Guttman 1945) dont elles représentent en fait une estimation conservatrice, car elles n'ont pas été calculées en considérant la longueur totale du test.

TABLEAU 3.7
Corrélations entre items pairs et impairs

Items	# 1 COMPR		#2 ENAPP		#3 TROUS	
	Impairs	Pairs	Impairs	Pairs	Impairs	Pairs
#1 Impairs	1.000	0.996	0.857	0.857	0.873	0.873
#1 Pairs	0.996	1.000	0.853	0.853	0.873	0.873
#2 Impairs	0.857	0.853	1.000	0.992	0.809	0.811
#2 Pairs	0.857	0.853	0.992	1.000	0.805	0.808
#3 Impairs	0.873	0.873	0.809	0.805	1.000	1.000
#3 Pairs	0.873	0.873	0.811	0.808	0.995	1.000

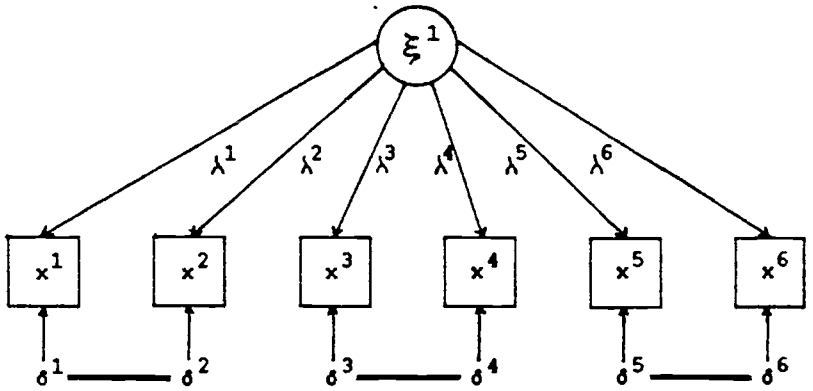
Comme toutes ces valeurs dépassent les coefficients KR-20 que nous avons calculés, on peut penser qu'un facteur externe, tel la fatigue, a pu affecter le calcul de la fiabilité selon la formule de Kuder et Richardson. Il apparaît donc justifié de postuler le parallélisme entre les deux parties de chaque sous-test puisque les deux parties semblent mesurer la ou les même(s) habileté(s).

Selon la suggestion de McDonald (1985), nous avons d'abord cherché à vérifier le modèle qui nous semblait le plus probable et qui en même temps était sans doute le plus simple. Il s'agit d'un modèle d'analyse factorielle c'est-à-dire un modèle qui ne comporte pas d'équation structurale et un seul modèle de mesure définit par la simple équation:

$$x = \Lambda x \xi + \delta \qquad (3.9)$$

La première hypothèse que nous posons restreint le modèle à un seul facteur, soit une seule variable ξ. La figure 3.5 montre le schéma correspondant au modèle. La direction des flèches indique que les variables observables sont affectées par la variable latente et par l'erreur de mesure tandis que les lignes doubles montrent que les éléments ne sont pas significativement différents (parallélisme). En soumettant cette première hypothèse au programme *LISREL*, nous nous sommes aperçu qu'elle ignorait une bonne portion de la variance et de la covariance et qu'une fois le premier sous-test ajusté au modèle, il restait pour les deux autres sous-tests, des valeurs résiduelles importantes qui rendaient ce modèle peu adéquat.

FIGURE 3.5
Schéma d'un modèle de tests parallèles avec facteur unique

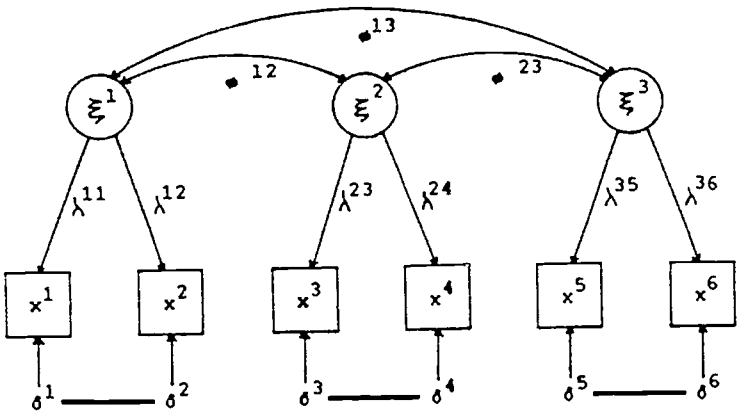


Avant de passer au modèle congénérique, nous avons voulu vérifier une seconde hypothèse qui faisait aussi appel à un modèle de l'analyse factorielle. Plutôt que de contraindre les données à facteur unique nous avons postulé trois facteurs fortement corrélés. Dans la figure 3.6, qui illustre ce modèle, les flèches bi-directionnelles indiquent une influence réciproque entre les variables latentes. L'examen des estimations de *LISREL* montre que ce modèle s'avère très juste. De fait, on peut se demander s'il ne s'agit pas d'une solution trop parfaite pour être généralisable à d'autres tests de langue. Il semblait donc inutile de faire intervenir d'autres variables à ce moment-ci. D'autre part, on s'aperçoit que les corrélations entre les facteurs sont très élevées puisqu'elles varient entre .81 et .87.

La solution oblique est juste mais elle n'est pas tout à fait satisfaisante dans la mesure où elle laisse inexpliquée la variance commune entre les facteurs. Un modèle congénérique aurait peut-être pu contribuer à préciser davantage la relation entre les éléments mais nous avons été incapable de trouver une solution adéquate en utilisant le modèle congénérique. L'addition d'autres mesures dans la matrice aurait pu peut-être servir à mettre au point un tel modèle mais il s'agit là d'une tâche qui dépasse les

buts de la présente recherche. Par ailleurs, il faut bien se rendre compte que ce type de solution où les facteurs spécifiques à une tâche langagière sont en corrélation avec les facteurs reliés à d'autres tâches est sans doute ce à quoi il faut s'attendre dans les tests de langue. Il semble bien qu'en ce qui concerne la performance communicative, peu importe les distinctions qu'on aura établies, les aspects dégagés seront toujours imbriqués les uns aux autres.

FIGURE 3.6
Schéma d'un modèle de tests parallèles avec trois facteurs



Dans le cas du test que nous mettons au point, cette analyse se révèle utile à deux points de vue. D'une part, elle nous assure que chaque sous-test présente une excellente cohérence interne. Il faut noter que cela ne garantit pas l'unidimensionalité de chaque sous-test. Néanmoins, on peut penser qu'à la suite d'une étude approfondie des corrélations entre les items, où l'on arriverait à neutraliser l'effet du facteur de difficulté, on pourrait expliquer la variance à l'aide d'un facteur dominant. Compte tenu de l'objectif de notre démarche, une telle analyse ne nous apparaissait pas nécessaire car des recherches sur les tests de langue, dans le cadre de la théorie du trait latent, ont déjà démontré que des tests qui présentent une bonne cohérence interne permettent une paramétrisation fiable (Henning 1984, Henning *et al.* 1985, Cook *et al.* 1988, Davidson 1988). Il apparaît également que, dans notre cas, chaque sous-test

comporte une variance spécifique de sorte qu'il serait imprudent d'intégrer tous les items dans une seule banque. Il convient donc de procéder à une calibration indépendante pour chaque sous-test. D'autre part, le fait que les scores des sous-tests de même que les facteurs qui les sous-tendent soient corrélés démontre qu'il est raisonnable d'espérer obtenir un indice de la maîtrise générale en combinant les résultats de chaque partie du test.

3.2.2.2.3 Corrélations avec d'autres mesures

Nous avons retenu 51 étudiants de l'Université Laurentienne pour lesquels nous avions plusieurs mesures valables à mettre en corrélation. Outre les trois parties du test expérimental, nous avions à notre disposition les résultats du test Laval et l'assignation définitive à un des cinq groupes-classes du programme. De plus, lors du test, on demandait aux étudiants d'identifier le niveau auquel ils estimaient appartenir parmi les sept niveaux que nous avons reconnus; on leur demandait aussi d'indiquer le nombre d'années d'apprentissage du français à l'école secondaire et/ou dans un établissement post-secondaire. Toutes ces mesures peuvent être considérées comme des mesures à intervalles et peuvent donc être comparées entre elles. Le tableau 3.8 donne les moyennes et les écarts types calculés pour chacune de ces mesures. Le test expérimental a été administré à la deuxième semaine d'une session intensive de six semaines. Les variances (comme les coefficients de corrélation, du reste) sont en général légèrement moins élevées que celles de l'ensemble de l'échantillon d'analyse.

TABLEAU 3.8
Moyennes et écarts types des mesures concurrentes de la version 2

	Moyenne	Écart-type
LAVAL: Test Laval (trois parties)	62.33	29.98
COMPR: #1 Compréhension	29.96	10.87
ENAPP: #2 Énoncé approprié	29.75	7.83
TROUS: #3 Phrases à trou	27.59	8.34
GROUPE: Assignation (5 groupes)	3.09	1.41
AUTO: Auto-évaluation(7 catégories)	3.51	1.65
SCOL: Scolarité (Nombre d'années)	3.65	2.31

Le classement initial des étudiants s'est effectué avec le *test Laval*. Ce test comprend trois sections composées de questions à choix multiple: phonétique (30 items), grammaire (75 items) et vocabulaire (50 items). Bien que basé sur une approche à éléments discrets qui a de moins en moins la faveur des enseignants, la commodité du *test Laval* explique pourquoi il continue d'être aussi populaire dans des programmes comme celui de la Laurentienne. Au début de la session, avant que nous n'administrions le test expérimental, le directeur pédagogique avait dû changer 24% des étudiants de groupe, le plus souvent à cause d'erreurs de classement. Les assignations dont nous tenons compte (GROUPE) réflètent le classement après les changements mais il faut néanmoins s'attendre à ce que ces assignations soient assez fortement corrélés avec le *test Laval*. Ainsi que le montre le tableau 3.9, c'est d'ailleurs avec les assignations que le score du *test Laval* présente le coefficient de corrélation le plus élevé. Malgré le nombre restreint de sujets, toutes les corrélations sont significatives ($p < .001$). La corrélation la plus élevée s'observe entre l'auto-évaluation (AUTO) et l'assignation finale (GROUPE). Cela s'explique autant par le nombre limité de niveaux (5 et 7) que par le fait que les étudiants, une fois les changements de groupe effectués, estiment que le groupe auquel ils appartiennent est celui qui correspond effectivement à leur niveau.

TABLEAU 3.9

Corrélations entre les mesures concurrentes

	LAVAL	COMPR	ENAPP	TROUS	AUTO	SCOL	GROUPE
LAVAL	1.000	.746	.610	.649	.753	.559	.879
COMPR	.746	1.000	.819	.854	.829	.618	.868
ENAPP	.610	.819	1.000	.780	.750	.658	.772
TROUS	.650	.854	.780	1.000	.866	.675	.825
AUTO	.753	.829	.750	.865	1.000	.644	.916
SCOL	.559	.618	.658	.675	.644	1.000	.694
GROUPE	.879	.868	.772	.825	.916	.694	1.000

Ce qui étonne toutefois, c'est que la corrélation entre l'auto-évaluation et les scores du *test Laval*, celui-là même qui a servi au classement initial, soit si faible. De fait, de ce point de vue, ce sont

les scores des sous-tests #1 (COMPR) et #3 (TROUS) qui seraient les meilleurs prédicteurs puisqu'ils montrent des corrélations élevées à la fois avec l'auto-évaluation et l'assignation. Le sous-test #2 (ENAPP), de par son contenu et une fiabilité plus faible, engendre des corrélations relativement peu élevées. Enfin, le nombre d'années d'apprentissage du français à l'école (SCOL) semble un prédicteur plutôt médiocre de sorte qu'il faudrait l'utiliser avec prudence lors de l'estimation du niveau général d'un étudiant.

Afin de compléter ces premières études sur la validité du test, nous avons voulu voir comment, à l'intérieur d'un groupe d'étudiants, le score au test se comparait avec l'évaluation du professeur. Celà nous semblait d'autant plus important que c'est généralement le jugement des professeurs qui sert à rectifier les erreurs de classement. C'est pourquoi, lorsque c'était possible, nous remettions aux professeurs une feuille sur laquelle ils devaient dresser la liste de leurs étudiants en les ordonnant à partir du moins avancé jusqu'au plus avancé. On ne leur fournissait pas de critères particuliers, mais on leur demandait d'assigner un rang en prenant en considération autant la maîtrise générale que les objectifs de leur cours et ce après au moins une semaine de cours (minimum de 25 heures). Comme les groupes dépassent rarement une quinzaine d'apprenants, l'exercice est tout à fait réalisable et plusieurs professeurs s'y sont prêtés.

Pourtant, il faut toutefois s'attendre à des corrélations ni très élevées, ni très significatives. D'abord, on a observé que si les professeurs s'entendent pour remettre en question un instrument de classement, l'unanimité disparaît quand il s'agit de s'entendre sur ce qui est prioritaire et ce qui l'est moins (Laurier 1984). En faisant l'analyse de la validité concurrente de tests de compréhension auditive, Groot (1975) a observé des écarts importants entre les jugements des enseignants. Alderson (1990) signale également que les jugements d'«experts» sont souvent contradictoires. Pour cette raison, il ne faut pas s'étonner qu'il y ait des différences notables entre les coefficients de chaque groupe et qu'on retrouve même des corrélations négatives. Ensuite, on effectue des distinctions à l'intérieur d'une gamme très étroite d'habileté, une gamme souvent même

plus étroite que l'intervalle de confiance défini par l'erreur de mesure du test. Cela explique en partie le fait que les coefficients ne dépassent guère .7. Enfin, on établit des corrélations à partir d'un nombre restreint de sujets (environ 15), ce qui a pour effet de diminuer considérablement les probabilités de trouver des corrélations significatives. De fait, il n'est pas du tout certain que les corrélations de rang permettent de bien saisir l'accord (ou le désaccord) entre le test et le jugement du professeur. Il serait sans doute souhaitable, dans une étude ultérieure, de compléter avec une procédure similaire à celle de Magnan (1987) pour l'entrevue et de calculer le coefficient *kappa* de Cohen. Berry et Mielke (1988) ont en effet démontré la robustesse de cet indice et son application avec des mesures ordinales.

Ces corrélations ont été établies en comparant le rang assigné par le professeur et celui qu'on obtenait à partir du score brut à chaque sous-test. Toutefois, afin d'obtenir des indices plus précis, il nous semblait approprié de considérer plutôt la moyenne des corrélations pour chaque partie. Afin d'éviter une distorsion des indices à la suite des opérations arithmétiques qu'implique le calcul des moyennes, on a appliqué la transformation du Z de Fisher, calculé les moyennes puis reconverti. On obtient ainsi une corrélation de .44 pour le sous-test de compréhension, .33 pour le choix de l'énoncé approprié et .37 pour les phrases à trou. Compte tenu des facteurs que nous avons signalés plus haut, une corrélation supérieure à .4 doit être interprétée comme l'indice d'une bonne correspondance entre les deux mesures.

On peut donc conclure avec ces premières données sur la validité du test que le premier sous-test s'avère le plus valide. Il s'agit là de résultats provisoires car, comme le font remarquer Allen et Yen (1979:108), l'analyse de la validité d'un test est un travail à long terme qui progresse à mesure que s'accumulent les résultats d'épreuves concurrentes et les observations des utilisateurs.

3.2.2.3 *Analyse des items*

En consultant la distribution des réponses que fournit CORREC et les coefficients de probabilité, on note que la représen-

tation de chaque niveau s'est améliorée car il y a nettement plus d'items difficiles et un peu plus d'items faciles. Toutefois l'item 22 du deuxième sous-test (P = .06) et l'item 35 du troisième sous-test (P =.09) sont nettement trop difficiles et devront sans doute être retirés.

Comme pour la version pré-expérimentale, nous avons soumis les résultats de l'échantillon d'analyse au programme *LERTAP* afin d'évaluer la discrimination de chaque item. Ce qui ressort en premier lieu, c'est que la partie de compréhension contient d'excellents items du point de vue de leur discrimination alors que le choix de l'énoncé approprié présente plusieurs items médiocres. Ainsi le sous-test #1 regroupe pas moins de 14 items dont le coefficient de corrélation point-bisérielle dépasse .5 et le coefficient de corrélation bisérielle dépasse .75. Parmi ceux-ci, les items 24, 25, 27, 36, 39 et 50 ont des corrélations point-bisérielles et bisérielles supérieures à .6 et .8, respectivement. On ne peut en dire autant du sous-test #2 où la norme .5 /.75 n'est dépassé, à peine, que par les items 34 et 40; les items 5 et 19 s'approchent de ce seuil. Les items de la partie des phrases lacunaires sont statistiquement assez valables bien que la grande difficulté de certains affecte probablement leur discrimination. On compte 9 items qui dépassent la norme des .5 /.75; les items 6, 20 et 42 présentent même des coefficients supérieurs à .6 et .8.

Nous avons aussi cherché à identifier les items qui montraient des caractéristiques statistiques moins reluisantes. Afin de compléter nos données sur les items douteux, nous avons soumis les réponses de l'échantillon d'analyse au programme *BICAL*, non pas pour la paramétrisation mais pour tenir compte des indices d'adéquation au modèle de Rasch que fournit ce logiciel. À ce moment, il ne s'agissait plus d'éliminer des items ou d'en modifier car nous comptions effectuer une première calibration avec un modèle à trois paramètres en soumettant les 50 items de chaque sous-test. Toutefois, il était important de réunir des données pour corroborer les résultats de la calibration advenant l'élimination de mauvais items lors de la calibration finale. D'autre part, nous voulions constituer une liste d'items susceptibles de ne pas appa-

raître dans les deux versions «papier-crayon» du test. Le tableau 3.10 montre les indices dont nous avons tenu compte pour les items douteux du premier sous-test.

TABLEAU 3.10
Items douteux du sous-test #1

Item #	SCALE Indice	LERTAP r_{pbis}	LERTAP r_{bis}	BICAL Disc.	BICAL Test T	CORREC Prob.
10	.93	.28	.48	.66	.10	.91
11	.72	.36	.46	.72	2.85	.40
13	.72	.46	.59	.67	3.76	.40
21	.70	.39	.50	.49	5.04	.35
30	.67	.45	.57	.53	5.08	.36
31	.67	.35	.44	.49	5.68	.45
40	.69	.41	.51	.77	2.49	.46
47	.83	.29	.41	.56	2.49	.78
49	.63	.33	.41	.31	7.31	.47

Les chiffres soulignés signalent les problèmes les plus sérieux. On remarque que dans le sous-test #1, seulement deux items, le 10 et le 49, présentaient des corrélations point-bisérielles inférieures à .3. Par contre, comme en fait foi le tableau 3.11, nous en avons identifié 15 au sous-test #2. Les items 11, 21, 22, 29 et 31 semblaient discriminer particulièrement mal.

TABLEAU 3.11
Items douteux du sous-test #2

Item #	Indice SCALE	LERTAP r_{pbis}	LERTAP r_{bis}	BICAL Disc.	BICAL Test T	CORREC Prob.
1	.70	.27	.34	.85	.78	.35
7	.76	.18	.25	.60	1.73	.79
8	.69	.17	.23	.62	1.74	.34
11	.69	.11	.15	.49	2.08	.30
21	.66	.13	.17	.17	4.22	.31
22	.91	0	0	.68	-.12	.06
26	.85	.24	.37	.82	-.02	.87
27	.66	.23	.29	.76	2.28	.45
28	.70	.29	.37	1.04	-.36	.40
29	.67	.11	.15	.45	3.86	.70
31	.64	.09	.12	.16	4.56	.67
35	.67	.25	.31	.61	3.25	.55
37	.71	.29	.30	.52	2.35	.65
41	.68	.31	.39	.81	1.30	.38
44	.67	.27	.34	.74	1.72	.47
46	.66	.34	.43	.90	.34	.50
47	.65	.22	.28	.62	3.25	.43

TABLEAU 3.12
Items douteux du sous-test #3

Item #	Indice SCALE	LERTAP r^{pbis}	LERTAP r^{bis}	BICAL Disc.	BICAL Test T	CORREC Prob.
2	.69	.31	.39	.73	2.75	.45
3	.71	.33	.42	.73	2.53	.45
7	.68	.40	.51	.54	4.53	.47
8	.69	.42	.53	.75	2.43	.43
13	.69	.40	.50	.61	3.27	.55
16	.86	.17	.28	.83	-.08	.11
17	.54	-.01	.01	-.32	11.81	.43
35	.86	-.09	-.15	.44	1.21	.09
37	.79	.26	.38	.77	.98	.19
39	.72	.35	.46	.63	3.11	.29
48	.72	.07	.11	.31	3.46	.18

Enfin, nous n'avons trouvé que 4 coefficients de corrélation point-bisérielle inférieurs à .3 au sous-test #3 (tableau 3.12); il faut noter cependant que deux d'entre eux, le 17 et le 35, montraient une corrélation légèrement négative.

3.2.2.4 *Calibration des items*

L'opération de calibration des items revient à faire correspondre une courbe logistique à un ensemble de points obtenus à partir des données. Les paramètres ainsi obtenus déterminent l'équation reliant l'habileté d'un sujet avec le score obtenu à un item.

3.2.2.4.1 La procédure de calibration

À l'instar de Yen (1983), nous comptions utiliser pour la calibration définitive un modèle à trois paramètres. Lorsque nous avons lancé la présente recherche, nous avions prévu utiliser le programme le plus populaire et le plus accessible à cette époque pour la calibration à trois paramètres, le programme *LOGIST* (Wingersky *et al.* 1982). Ce logiciel mis au point par le *Educational Testing Service*, était installé sur l'ordinateur VAX de l'Institut d'Études pégagogiques de l'Ontario. Comme il fonctionne sur un

ordinateur central et que son exécution est assez longue, il est coûteux à utiliser. De plus, il demande une certaine familiarisation de l'usager qui doit intervenir pour spécifier le cadre de la paramétrisation. Ces inconvénients font toutefois partie du prix à payer pour un instrument souple et puissant.

Le programme fait l'estimation des paramètres des items et de l'habileté des sujets en suivant une procédure par maximum de vraisemblance. Il distingue entre les items dont les réponses ont été omises ou annulées (considérées comme incorrectes) et ceux auxquels le sujet n'a pas eu le temps de répondre (non retenus pour la calibration). L'usager doit s'assurer d'avoir retiré les sujets dont le score était nul ou parfait; de même les items dont les réponses sont toutes incorrectes ou correctes doivent être éliminés. La calibration se fait en quatre étapes. À la première et à la troisième étape on fait l'estimation des valeurs qui doivent être placées sur une échelle commune (l'origine est indéterminée): l'habileté des sujets et la difficulté des items. À la deuxième et à la troisième étape on fixe l'habileté et on estime les trois paramètres des items. La procédure est itérative c'est-à-dire qu'on répète l'opération à partir des dernières valeurs obtenues et ce, jusqu'à ce que le changement d'une itération à l'autre soit inférieur à un seuil pré-déterminé (par l'usager ou par le programme).

Les essais que nous avons menés avec *LOGIST* à l'Institut d'études pédagogiques de l'Ontario se sont avérés fructueux avec l'échantillon d'analyse. Pourtant, il nous a été impossible d'obtenir des résultats avec l'échantillon de calibration; il semblait que lorsque l'échantillon devenait trop grand le programme perdait une partie de l'information. Nous avons donc renoncé à utiliser cette version de *LOGIST* d'autant plus que d'autres problèmes rapportés par Vetterli (1987) nous auraient probablement amené à éliminer plusieurs items que nous voulions conserver.

Nous nous sommes alors tournés vers un nouveau produit qui venait de faire son apparition sur le marché des logiciels pour micro-ordinateurs et qui était disponible à notre lieu de travail, le

programme *ASCAL* (Assessment System Corp. 1987)[7]. Ce programme fait partie d'une batterie de logiciels conçus pour l'élaboration de tests conventionnels ou adaptatifs. Par rapport à *LOGIST*, *ASCAL* gagne en simplicité d'utilisation ce qu'il perd en flexibilité. Cependant comme nous menions une paramétrisation standard, ce logiciel était parfaitement approprié. De plus, une étude de Vale et Gialluca (1988) a démontré la supériorité de *ASCAL* par rapport à *LOGIST* particulièrement pour l'estimation du facteur de hasard ou lorsque l'échantillon regroupe moins d'un millier de sujets. De fait, l'algorithme de *ASCAL* s'apparente à celui de *LOGIST*, mais il intègre des principes d'analyse bayésienne. Ainsi, on amorce l'estimation finale à partir d'estimations initiales elles-mêmes obtenues en utilisant des procédures heuristiques traditionnelles: on suppose une distribution normale de a et b et on fixe c à la réciproque du nombre d'options de réponse. Ces estimations initiales servent à obtenir une première estimation de l'habileté des sujets pour laquelle on postule une distribution normale. Cette distribution est divisée en 20 groupes qui serviront au calcul final de a et c selon une procédure bayésienne où ces valeurs ont, au départ, une distribution beta. Quant au paramètre c, il est obtenu par une procédure utilisant le maximum de vraisemblance. La division en 20 groupes sert également au calcul d'un indice de l'adéquation de chaque item: on effectue le test du chi-carré sur les valeurs résiduelles dans chaque groupe. Cet indice doit être interprété en regard des indices obtenus aux autres items, une valeur beaucoup plus grande indiquant l'inadéquation de l'item. La procédure est itérative et s'arrête quand le changement d'une itération à l'autre devient marginal ou qu'on a atteint le nombre maximal d'itérations. Il faut noter que comme le programme fonctionne avec un micro-ordinateur, une telle calibration peut prendre plusieurs heures.

3.2.2.4.2 La première calibration

La structure factorielle que nous avions dégagée avec le programme *LISREL* nous incitait à la prudence de sorte que chaque

[7] Nous remercions Stan Jones de l'Université Carleton pour nous avoir permis d'utiliser le système *MicroCAT* installé au département de linguistique.

sous-test a été calibré indépendamment. Nous avons, dans un premier temps, soumis tous les items pour tous les sujets qui avaient été conservés. Nous espérions ici éviter de retirer des items qui malgré un piètre rendement en regard des statistiques traditionnelles auraient pu être jugés beaucoup plus favorablement dans le cadre d'une analyse du trait latent. La fonction de la première calibration était donc de servir d'analyse confirmatoire en vue de l'élimination des items moins efficaces. Nous nous sommes alors inspiré des lignes directrices établies par Urry (1977) pour fonder nos jugements sur la valeur de chaque item dans le cadre de la théorie du trait latent.

À la suite de l'analyse des items et de la première calibration, nous avons décidé de retirer quatre items du premier sous-test: les items 10, 31, 47 et 49. Les items 10 (mode d'emploi d'un café instantané), 47 (une carte postale à un ami) nous semblaient fort intéressants du point de vue de leur contenu, mais n'ont pu être conservés en raison de leurs pauvres caractéristiques psychométriques. Au deuxième sous-test, nous avons retiré une douzaine d'items (les items 7, 8, 11, 21, 22, 26, 27, 29, 31, 35, 37 et 47). Certains d'entre eux référaient à des notions ou des fonctions que nous aurions bien aimé voir figurer dans le contenu du test: suggérer (8 et 21), s'excuser (11 et 47), demander du feu (27) et laisser un message au téléphone (7). Nous nous sommes rendu compte combien il était difficile de créer de bons items pour mesurer ces fonctions et ces notions. Enfin, au troisième sous-test, nous sommes départis de huit items (les items 7, 9, 16, 17, 35, 37, 47 et 48) dont un d'entre eux à regret. En effet, l'item 47 évaluait un élément qui nous semblait devoir être vérifié au niveau avancé: l'usage des pronoms relatifs.

3.2.2.4.3 La deuxième calibration

Nous avons procédé à une autre calibration en excluant les items qui ne devaient pas être intégrés à la banque. De cette façon, on s'assurait que les paramètres soient le plus précis possible. Ce sont ces résultats que nous avons utilisés par la suite.

On peut voir en comparant les deux calibrations que, sans être tout à fait identiques, les paramètres n'ont pas changé sensiblement. Ainsi qu'il fallait s'y attendre, c'est le sous-test de compréhension qui présente les coefficients de discrimination les plus élevés (paramètre *a*) et le choix de l'énoncé approprié qui présente les coefficients de discrimination les plus bas. Dans les trois tests, on observe une distribution du niveau de difficulté qui devrait convenir aux fins que nous poursuivons avec ce test. De plus, l'indice du chi-carré témoigne du fait que tous les items retenus lors de cette dernière calibration cadrent bien avec le modèle à trois paramètres. Enfin, il faut noter qu'au terme de la calibration tous les items avaient convergé; de fait, les seuls items qui ont posé des problèmes au cours de la seconde calibration sont les deux derniers items du sous-test #3.

3.2.3 *La mise au point des versions équivalentes*

Il est clair que la version 2.2 du test posait des problèmes d'administration importants. Pour beaucoup d'établissements, il n'est pas possible d'utiliser un test qui dure deux heures et demie, peu importe les considérations psychométriques en jeu. Par ailleurs, il est certain qu'après une heure, il s'installe chez l'étudiant une certaine lassitude d'autant plus qu'avec seulement trois parties, l'épreuve peut devenir monotone. En élaborant une version expérimentale de 150 items, nous espérions non seulement conserver le nombre minimal d'items pour constituer une véritable banque, mais nous comptions également constituer deux versions «papier-crayon», plus courtes et équivalentes. En effet, il n'est pas rare qu'on ait besoin de plus d'une version d'un test, que ce soit pour alterner d'un semestre à l'autre ou même en vue d'une pratique aussi discutable que l'utilisation du test de classement comme pré-test puis comme post-test. Les deux versions finales du test «papier-crayon» (3.1 et 3.2) comprennent chacune 60 items soit 20 items par sous-test. Les items sont tous différents d'une version à l'autre sauf pour l'item 36 du premier sous-test de la version 2 et les items 4 et 23 du deuxième sous-test de la version 2. Ces items se retrouvent à la fois dans les versions 3.1 et 3.2 car ils s'avéraient particulièrement efficaces.

Les questionnaires se présentent sous la forme de fascicules (8½" x 11"), comprenant huit pages brochées. La couleur de la page frontispice identifie la version: bleu pour 3.1 et vert pour 3.2. Une grille de correction en acétate permet de corriger rapidement les feuilles de réponses.

3.2.3.1 *Le parallélisme*

En étudiant la structure factorielle du test, nous avions séparé pour chaque sous-test les items pairs des items impairs. Nous avions été surpris de constater que cette division laissait voir une cohérence interne remarquable. Il nous semblait donc logique d'utiliser cette division comme point de départ pour constituer les deux versions du test. En effet, les corrélations observées nous permettaient d'espérer en arriver non seulement à des formes équivalentes mais mieux encore, à des formes parallèles au sens où les définissent Lord et Novick (1968:chap 8): moyennes identiques et variances identiques.

Nous avons alors tenu compte du contenu des items non pas dans l'intention d'y retrouver à tout prix les mêmes éléments discrets d'une version à l'autre mais simplement pour éviter des duplications et nous assurer d'un certain équilibre entre les versions. Ainsi dans la première partie, deux items impliquaient la notion de comparaison (9 et 42) et dans la troisième partie, deux items (21 et 45) concernaient la notion d'antériorité dans le passé; nous avons donc partagé ces items entre les deux versions. Par contre, nous avons dû, pour maintenir le parallélisme des sous-tests #2, placer les deux items reliés à la fonction «féliciter» dans la même version; nous ne croyons pas que cela affecte pour autant la validité du test dans la mesure où celui-ci ne repose pas sur un découpage strict d'un contenu notionnel, fonctionnel ou structural pré-établi. Nous avons cherché à répartir entre les deux versions, les nombreux items de la première partie impliquant des relations temporelles, les items de la deuxième partie reliés à la fonction «demander un service» ou ceux de la troisième partie mesurant uniquement les connaissances lexicales. Dans tous les cas, il s'agissait d'une première tentative et plusieurs substitutions ont suivi dans le but d'assurer l'égalité des moyennes

et des variances. Chaque sous-test a été traité séparément et chaque modification dans la sélection des items était évaluée à l'aide du programme STATEST ou du programme TESTAT, deux logiciels dérivés de notre programme CORREC. On obtenait alors des statistiques générales sur les deux versions du sous-test de sorte qu'il était possible de vérifier l'égalité des moyennes et des variances.

Toutefois, le fait de trouver des moyennes et des variances égales ne garantit pas que les scores seront comparables pour chaque niveau d'habileté. Encore faut-il uniformiser les échelles[8]. Une fois la paramétrisation des items de la banque complétée, la théorie du trait latent offre une solution attrayante tant par sa simplicité que par le fait qu'elle utilise toute l'information disponible. Ainsi, plutôt que d'utiliser les techniques traditionnelles d'équivalence linéaire ou d'équivalence équipercentile (Angoff 1982), on peut comparer les courbes d'information des deux versions. Lorsque ces courbes sont identiques, on considère que les deux tests mesurent le trait de la même manière et que les scores sont donc comparables. Samejima (1977) fait une distinction entre le «parallélisme fort» et «parallélisme faible». Dans le premier cas, les statistiques des tests (moyennes, variances et corrélations) sont identiques et il y a une correspondance directe entre chaque item. Sans rencontrer toutes ces exigences, on peut viser le parallélisme «faible» en s'assurant simplement que les courbes d'information suivent le même tracé. Dans notre cas, on peut parler de «parallélisme moyen» puisque sans chercher une correspondance systématique entre les items de chaque version, nous avons essayé de concilier l'exigence de similarité des courbes d'information avec celle de l'égalité des variances et des moyennes. Il faut noter de plus que lorsque les courbes d'information coïncident, et que le nombre d'items est constant, on peut comparer directement les scores sans devoir les interpréter avec une échelle de conversion (Cook et Eignor 1983, Hambleton et Swaminathan 1985:chap 10). À l'aide du programme TICC, un programme complémentaire au système de gestion de la banque d'items, il nous était possible, pour chaque

[8] Il s'agit d'une opération que l'on nomme en anglais *scaling*, terme pour lequel nous n'avons pas trouvé de traduction satisfaisante.

nouvelle combinaison, de calculer les points de la courbe d'information en fonction des niveaux d'habileté qui nous intéressaient.

Avec une méthode d'essai et d'erreur, où l'effet de chaque modification dans la sélection des items était vérifié tant au plan des moyennes et des variances qu'au plan des courbes d'information, nous en sommes venus à une sélection optimale pour chaque sous-test.

TABLEAU 3.13
Moyennes et écarts types des versions 3

		Moyenne	Écart-type
Version 3.1	#1 Compréhension	11.69	5.03
	#2 Énoncé approprié	12.92	3.98
	#3 Phrases à trou	11.43	4.46
	TEST COMPLET	36.04	12.53
Version 3.2	#1 Compréhension	11.83	5.18
	#2 Énoncé approprié	13.00	3.93
	#3 Phrases à trou	11.75	4.75
	TEST COMPLET	36.58	12.92

Le tableau 3.13 montre que les moyennes et les variances entre les sous-tests de chaque version varient peu. On note toutefois une différence d'un demi-point dans les moyennes de l'ensemble du test et une variance légèrement plus élevée pour la version 3.2. Quant à la fiabilité de l'ensemble du test, elle a forcément diminué du fait qu'on a conservé seulement 60 items mais elle demeure assez élevée, pour un test qui ne dure qu'une heure: on a calculé un indice KR-20 de .923 pour la version 3.1 et de .927 pour la version 3.2.

Comme le montrent les figures 3.7a et 3.7b, les courbes d'information des sous-tests sont assez semblables d'une version à l'autre. Les deux versions du premier sous-test (Compréhension) donnent un maximum d'information dans la zone -0.1 de l'échelle d'habileté (entre le niveau intermédiaire faible et le niveau intermédaire moyen). Malgré un indice de fiabilité légèrement supérieur, on voit que la version 3.2 de ce sous-test donne généralement un peu moins d'information mais l'écart entre les courbes reste tout de même minime et constant. Étant donné que nous anticipions une marge d'erreur plus importante au deuxième sous-test (Énoncé

approprié), il n'est pas surprenant que les courbes d'information soient moins élevées. Ce sous-test donne un maximum d'information dans la zone 0.4 soit à la frontière entre le niveau intermédiaire fort et le niveau avancé. C'est d'ailleurs dans cette zone que la différence entre la quantité d'information obtenue avec chaque version est la plus importante. En dehors de cette zone, même si la version 3.1 semble supérieure, l'écart est négligeable. Enfin, le dernier sous-test (Phrases à trou) fournit le maximum d'information dans la même zone que le deuxième mais avec une marge d'erreur plus réduite. À partir du sommet de la courbe jusqu'au niveau le plus avancé, l'écart entre les deux versions s'agrandit mais cette fois en faveur de la version 3.2 plutôt que de la version 3.1. Ainsi dans l'ensemble, on obtient de l'information pour tous les niveaux d'habileté mais la distribution n'est pas rectangulaire. Il serait utopique d'ailleurs de viser un tel objectif puisqu'on tend généralement à éliminer les items très difficiles ou très faciles. On peut même alléguer qu'une distribution rectangulaire n'est pas nécessairement souhaitable car c'est généralement au niveau intermédiaire que les professeurs ressentent le besoin d'étayer leur jugement avec des mesures plus objectives.

FIGURE 3.7a

Courbes d'information de la version 3.1

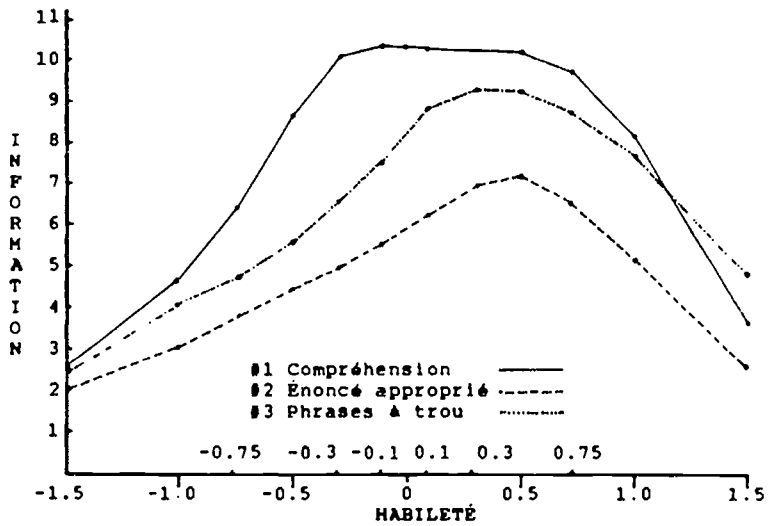
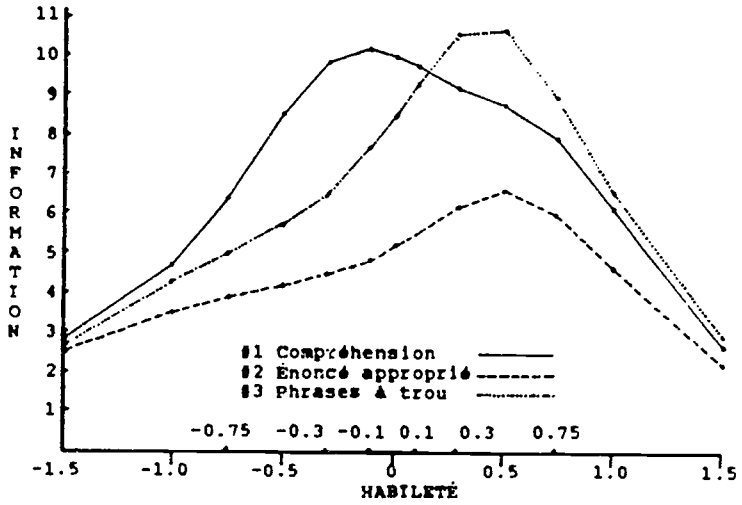


FIGURE 3.7b
Courbes d'information de la version 3.2



Une fois la sélection complétée, nous avons ordonné les items de chaque sous-test, à partir de l'item le plus facile jusqu'au plus difficile. Cette réorganisation suit la suggestion de Spolsky *et al.* (1972) et évite à l'étudiant débutant de se voir confronté à des questions trop difficiles. Ainsi, beaucoup de ces débutants choisiront de ne pas répondre dès qu'ils auront l'impression de deviner la réponse. D'autre part, on a placé, à la toute fin de chaque partie, un item de difficulté moyenne de sorte que le sujet quitte l'épreuve ou la section sans une impression d'échec. On reconnaît dans cette progression, les principes méthodologiques de l'entrevue pour l'évaluation de l'expression orale. Il est certain que l'ordre des items affecte la distribution des scores (Hambleton et Traub 1974) de sorte qu'il faudra probablement, après un certain temps, réajuster la répartition des niveaux.

3.2.3.2 *La répartition des niveaux*

Étant donné qu'un test de classement vise le plus souvent à attribuer un niveau plutôt qu'à assigner un score, il était important d'établir un barème permettant de faire correspondre le score obtenu à la version 3.1 ou 3.2 avec un des sept niveaux que nous comptions

distinguer. Nous avons donc, dans un premier temps, divisé la courbe normale en sept tranches à l'intérieur desquelles, en supposant une distribution tout à fait normale, on devait retrouver un nombre égal de sujets. La répartition s'établit selon la figure 3.8. À l'aide du programme TESTAT, nous avons ordonné les sujets de l'échantillon d'analyse selon leur score en ne retenant que les items figurant dans la version 3.1 puis en ne retenant que ceux figurant dans la version 3.2.

FIGURE 3.8

Répartition des niveaux dans la population

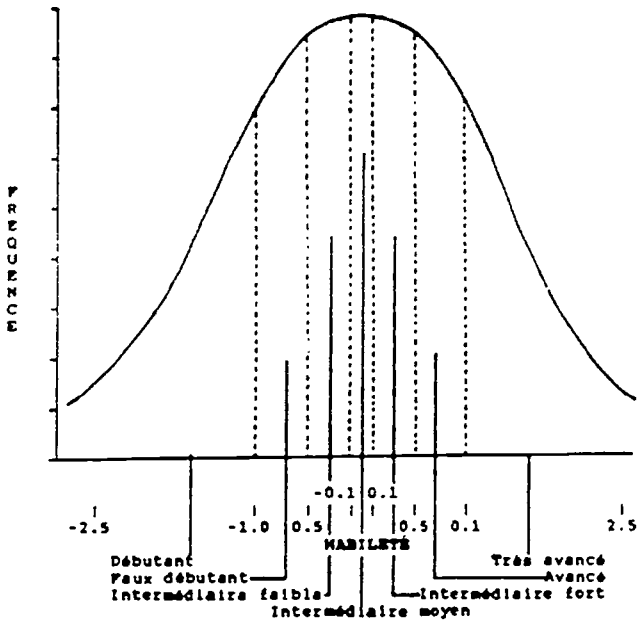


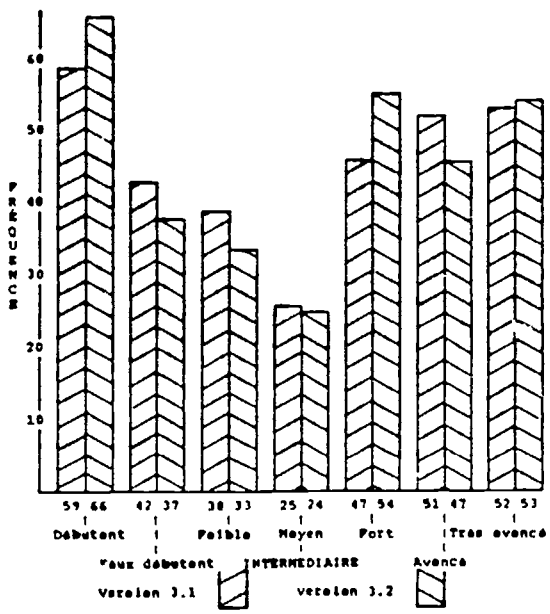
TABLEAU 3.14

Barème de correction des versions 3.1 et 3.2

Niveau	Z-Std	Score / 60
Débutants	Entre -3.0 et -1.0	De 0 à 23
Faux débutants	Entre -1.0 et -0.5	De 24 à 30
Intermédiaire faible	Entre -0.5 et -0.1	De 31 à 34
Intermédiaire moyen	Entre -0.1 et 0.1	De 35 à 37
Intermédiaire fort	Entre 0.1 et 0.5	De 38 à 43
Avancé	Entre 0.5 et 1.0	De 44 à 49
Très avancé	Entre 1.0 et 3.0	De 50 à 60

En considérant les scores *z* standardisés pour chaque version, nous avons cherché à établir une échelle qui puisse minimiser les écarts entre les résultats selon chaque version tout en respectant la répartition des niveaux que nous avions établie. On obtenait ainsi une échelle unique pour les deux versions (tableau 3.14). Bien que notre méthode introduise une source d'erreur supplémentaire, il nous semblait que le parallélisme des deux versions était tel que la commodité d'une échelle unique compensait largement une diminution probablement négligeable de la fiabilité. L'histogramme de la figure 3.9 permet de comparer la distribution des sujets de l'échantillon d'analyse selon leur score à la version 3.1 ou à la version 3.2. La somme des différences absolues entre les niveaux où chaque étudiant a été placé est de 30. Avec 314 sujets, on obtient donc un taux d'erreur de classement inférieur à 10%. Ce taux est tout à fait acceptable si l'on tient compte de la proportion normale de cas frontières, c'est-à-dire de sujets qui pourraient être placés dans n'importe lequel de deux niveaux adjacents.

FIGURE 3.9
Répartition des niveaux aux version 3.1 et 3.2



3.2.3.3 *Les corrélations*

3.2.3.3.1 Les corrélations entre les sous-tests

Afin de confirmer le parallélisme des versions, nous avons établi des corrélations entre les scores pour chaque sous-test et le score global des deux versions finales. Ces corrélations qu'on retrouve au tableau 3.15 tendent à démontrer que les deux versions sont parallèles.

TABLEAU 3.15
Corrélations entre les sous-tests 3.1 et 3.2

	Version 3.1				Version 3.2			
	#1	#2	#3	Total	#1	#2	#3	Total
3.1								
#1	1.000	.811	.821	.952	.911	.839	.819	.920
#2	.811	1.000	.748	.910	.794	.830	.775	.854
#3	.821	.748	1.000	.923	.815	.768	.844	.869
Total	.952	.910	.923	1.000	.908	.874	.875	.950
3.2								
#1	.911	.794	.815	.908	1.000	.812	.830	.952
#2	.839	.830	.768	.874	.812	1.000	.769	.911
#3	.819	.775	.844	.875	.830	.769	1.000	.933
Total	.920	.854	.869	.950	.952	.911	.933	1.000

Toutes les corrélations sont significatives ($p < .001$) et en examinant la matrice, on voit que les différences entre les coefficients de chaque quadrilatère sont attribuables aux variations d'un sous-test à l'autre plutôt qu'aux variations entre les versions. On peut s'étonner de ce que ces corrélations soient plus faibles que celles que nous avions calculées en comparant les items pairs avec les items impairs de la version 2. Cela tient sans doute au fait que les coefficients n'ont pas été corrigés pour tenir compte de la fiabilité. Or, avec 20 items au lieu de 25, il faut s'attendre à une certaine chute des coefficients. Par ailleurs, il n'est pas exclu que le facteur de fatigue ait été mieux neutralisé avec la division pair/impair ou que l'effet du hasard ait tout simplement donné des corrélations particulièrement élevées lors de cette première division.

3.2.3.3.2 Les corrélations avec d'autres mesures

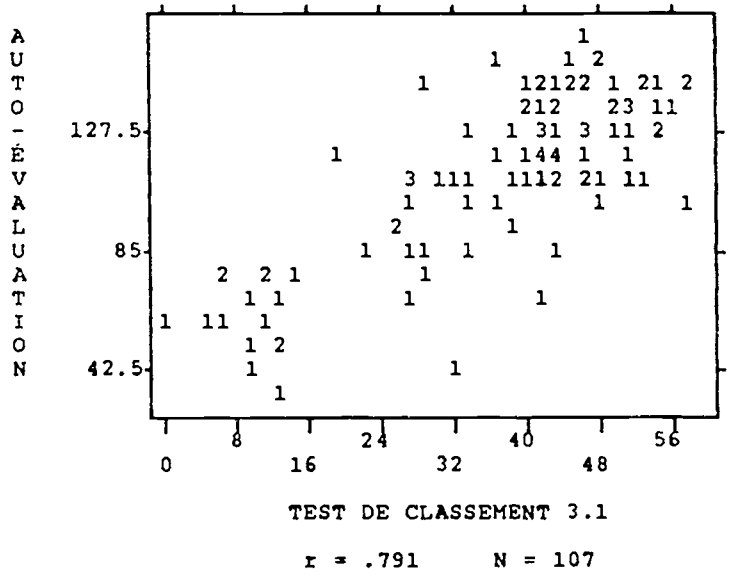
Si le test est utilisé par plusieurs établissements, nous aurons l'occasion, au fil des ans, d'accumuler des données qui nous

permettront d'une part, de recalibrer les items avec un échantillon plus large et d'autre part, de poursuivre la validation du test. Nous avons néanmoins recueilli certains renseignements à cet égard. Depuis deux ans, le Collège de Saint-Boniface utilise le test comme instrument de classement pour ses cours d'été et semble tout à fait satisfait du rendement du test. Au cours de l'année scolaire 1988-89, nous avons aussi utilisé la version 3.1 auprès d'étudiants inscrits à l'Université Carleton non seulement comme moyen de vérifier leur niveau mais également afin d'étudier la validité concurrente du test puisque d'autres mesures étaient disponibles.

Il nous semblait particulièrement intéressant de comparer les résultats d'une des versions finales du test avec un questionnaire d'auto-évaluation. On note en didactique des langues, une tendance à intégrer de plus en plus l'auto-évaluation à la mesure de la maîtrise (Jarmasz 1983, Lewkowicz et Moon 1985). Connors (1983) a trouvé que l'évaluation que font les apprenants de leur propre apprentissage concorde souvent avec celle de juges. Notre instrument d'auto-évaluation utilisait en grande partie les situations du questionnaire en usage depuis plusieurs années à l'Université d'Ottawa (Leblanc 1989). Il s'agit de demander à l'étudiant d'évaluer sur une échelle de fréquence à cinq catégories (à partir de «jamais» jusqu'à «toujours») comment il estime pouvoir accomplir certaines tâches dans la langue seconde. On peut, par exemple, demander à l'étudiant s'il peut suivre l'intrigue d'un film en français. Cette méthode se révèle habituellement assez efficace mais peu fiable dans certaines circonstances, notamment quand l'étudiant a de bonnes raisons de vouloir fausser son résultat. Comme le questionnaire de l'Université d'Ottawa n'évalue que les habiletés réceptives et ce, dans le cadre d'une institution bilingue, nous avons éliminé plusieurs questions que nous jugions peu pertinentes et nous avons ajouté une section sur la production. De plus, nous nous préoccupions surtout de voir si notre test pouvait mesurer les habiletés orales; nous avons donc éliminé toute référence à la langue écrite. Les deux sections du questionnaire ont été administrées à une centaine d'étudiants de l'Université Carleton; tous les niveaux étaient représentés et on avait prévenu les étudiants qu'ils ne risquaient pas de devoir changer de classe à la suite de ces tests.

En comparant les résultats de 107 étudiants au questionnaire d'auto-évaluation avec les résultats de la version 3.1, nous avons trouvé une corrélation de .79 entre les deux instruments. La corrélation est relativement élevée compte tenu des problèmes de fiabilité que nous avons constatés avec l'auto-évaluation. Le diagramme de dispersion de la figure 3.10 représente la relation entre les deux mesures.

FIGURE 3.10
Diagramme de dispersion de l'auto-évaluation et de la version 3.1



Le diagramme a été obtenu avec *SPSS-PC* (Norusis *et al.* 1988) et les chiffres représentent le nombres d'observations à un point donné. De façon générale, les corrélations entre chacune des parties des deux épreuves dépassent .75 et elles sont légèrement plus élevées avec la section «expression» de l'auto-évaluation.

On a aussi demandé aux mêmes étudiants de remplir un questionnaire sur leurs connaissances lexicales. Nous espérions en

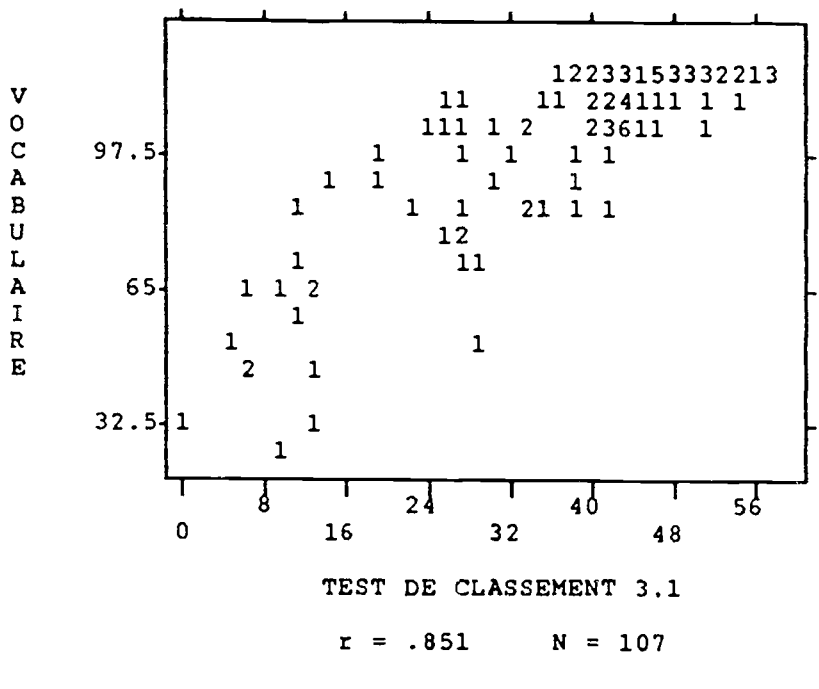
effet découvrir pourquoi les corrélations entre les sous-tests (ou entre les facteurs, si on se réfère aux résultats obtenus avec *LISREL*) étaient si élevées. Saville-Troike (1984) a mis en lumière le rôle du vocabulaire chez des étudiants devant fonctionner dans un contexte scolaire en langue seconde. Barnett (1986) a trouvé que le vocabulaire était un élément déterminant dans la lecture en langue seconde. Il nous semblait aussi que la connaissance du vocabulaire pouvait expliquer la correspondance entre les résultats de différentes épreuves. Afin de vérifier cette hypothèse, nous avons identifié le mot clé le plus important dans chaque item de la version 3.1. Ensuite, nous avons placé ces soixante mots dans un contexte d'une phrase où le mot conservait la même valeur sémantique que dans l'item original. On demandait aux étudiants de dire s'ils connaissaient le sens du mot peu importe le contexte utilisé (2 points), s'ils pouvaient en deviner le sens à partir du contexte fourni (un point) ou s'ils en ignoraient tout à fait le sens. Cette formule s'inspire des recherches de Meara et Buxton (1987) qui ont trouvé qu'une liste de mots à cocher était un instrument plus fiable que le traditionnel test lexical à choix multiple.

Nous avons trouvé, entre notre épreuve de vocabulaire et la version 3.1, une corrélation de .85 chez les 107 étudiants qui avaient fait les deux épreuves. La corrélation n'est pas aussi élevée que ce à quoi nous nous attendions mais l'examen du diagramme de dispersion (figure 3.11) est particulièrement révélateur. On constate que pour les étudiants plus avancés (40 et plus à la version 3.1), les scores du test de vocabulaire tendent à s'agglutiner autour des scores les plus élevés.

Ce plafonnement indiquerait que la mesure des connaissances lexicales pourrait être une bonne indication de la maîtrise de la langue seconde pour les niveaux débutants et intermédiaires mais qu'elle cesserait de l'être pour les plus avancés. Cette interprétation va dans le sens des recherches de Adams (1980) qui a trouvé que chez les débutants, le facteur prédominant était le vocabulaire mais que l'importance de ce facteur s'estompait quand les sujets étaient plus avancés. De même, Jochens et Montens (1988) attribuent la

réduction de la variance des tests de closure au niveau avancé, au fait que le vocabulaire le plus fréquent est déjà connu. Il nous semble donc que le lexique joue un rôle important surtout chez les débutants. Il s'agit d'une hypothèse de recherche qui mériterait d'être explorée plus à fond mais cette tâche dépasse les limites de la présente étude.

FIGURE 3.11
Diagramme de dispersion de l'épreuve de vocabulaire et de la version 3.1



Toujours pour nous en tenir aux objectifs initiaux de notre recherche, nous n'avons pas examiné en détail la matrice des corrélations obtenue avec les résultats des 86 sujets qui ont fait les trois épreuves (tableau 3.16). Toutes les corrélations sont significatives ($p < .001$). On observe que tous les coefficients impliquant le test de vocabulaire sont supérieurs à .8 et ce, même pour les deux sections de l'auto-évaluation.

TABLEAU 3.16
Corrélations entre l'auto-évaluation, la version 3.1 et l'épreuve de vocabulaire (N = 86)

	Auto-évaluation			Version 3.1				Vocabulaire
	Ecoute	Expr.	Total	#1	#2	#3	Total	
Ec	1.000	.886	.964	.760	.748	.695	.773	.835
Ex	.886	1.000	.976	.754	.758	.754	.792	.837
T	.964	.976	1.000	.779	.776	.749	.807	.861
#1	.760	.754	.779	1.000	.862	.870	.958	.849
#2	.748	.758	.776	.862	1.000	.862	.951	.845
#3	.695	.754	.749	.870	.862	1.000	.952	.816
T	.773	.792	.807	.958	.951	.952	1.000	.877
Vo	.835	.837	.861	.849	.845	.816	.877	1.000

Il ne fait donc aucun doute que le vocabulaire recouvre plusieurs aspects de l'utilisation de la langue seconde et qu'une étude approfondie de la composition factorielle de cette matrice pourrait être particulièrement révélatrice. Une telle étude pourrait notamment compléter les travaux de Harley *et al.* (1987) qui, dans une recherche auprès d'élèves d'immersion française, ont mis en lumière l'importance de la composante lexicale sans pouvoir toutefois déterminer s'il agissait d'un facteur autonome.

En complément à ces premières données sur la validité de la version 3, nous avons fait l'essai du test, comme outil de classement, en l'accompagnant d'un test mesurant spécifiquement la compréhension auditive. Ce test d'écoute consistait en quatre documents authentiques ou semi-authentiques enregistrés sur bande et suivis chacun de 15 questions de compréhension à choix multiples. Le test comprenait donc 60 items et deux versions différentes étaient disponibles (série 1 et série 2)[9]. Une expérimentation effectuée avec environ 250 boursiers a établi la fiabilité de la

[9] Ces deux versions du test de compréhension sont des versions expérimentales élaborées dans le cadre d'un contrat avec le Conseil des ministres de l'Éducation, l'organisme qui gère le Programme de bourses.

série 1 à .93 et celle de la série de la série à .87. On a administré la version 3.1 et la série 1 comme pré-test à une centaine d'étudiants du programme de bourses de Saint-Georges de Beauce (Université York). Après les six semaines du programme, on a administré, comme post-test, la version 3.2 et la série 2. Bien qu'on puisse penser que la compréhension de documents sonores fasse appel à une compétence qui n'est pas mesurée par le test que nous avons élaboré, les corrélations entre les deux instruments sont surprenantes: .82 pour le pré-test et .78 pour le post-test (tableau 3.17).

TABLEAU 3.17
Corrélations entre la version 3 et le test de compréhension auditive

	Pré-test		Post-test	
	Version 3.1	Série 1	Version 3.2	Série 2
Version 3.1	1.000	0.818	0.838	0.761
Série 1	0.818	1.000	0.763	0.843
Version 3.2	0.838	0.763	1.000	0.782
Série 2	0.761	0.843	0.782	1.000

Comme il s'agit de coefficients qui non pas été corrigés pour l'atténuation, il est permis de croire que la corrélation au post-test aurait sans doute égalé celle du post-test n'eût été de la fiabilité moindre de la série 2 et des mauvaises conditions qui prévalent généralement lors des post-tests. Par ailleurs, il ne faut pas s'étonner outre-mesure de trouver des coefficients aussi élevés car, rappelons-le, il s'agit d'un test de classement qui, par nature, est administré à une population où l'on trouve une gamme étendue de niveaux d'habilité. La variance tend donc à gonfler les coefficients de corrélation.

Bien sûr, il faudra faire suivre ces observations préliminaires d'expérimentations subséquentes afin de mieux déterminer la validité du test et de préciser davantage ce qu'il mesure. Comme le rappelle Cronbach (1971:452), c'est avec le temps que la validité du

construit se précisera: *Construct validation is therefore never completed. Construct validation is best seen as an ever-extending inquiry into the process that produces a high or low test score and into the other effects of those processes.* Néanmoins, les résultats que nous avons obtenus jusqu'ici nous permettent de croire que le test que nous avons élaboré, mesure vraiment une maîtrise générale du français. Il était donc tout à fait justifié de créer une version informatisée exploitant la banque d'items que nous avions mise sur pied.

④

MISE AU POINT DU DIDACTICIEL

Quand il s'agit de banques d'items, il faut distinguer deux catégories d'usagers. Premièrement, on trouve les responsables de l'évaluation et les concepteurs de programmes qui sont chargés de créer les items, de les expérimenter et finalement de les intégrer dans la banque. Ces spécialistes ont accès à la banque comme on accède à une base de données. Il faut donc prévoir à leur intention un système de développement. Deuxièmement, on trouve les sujets à qui on administre le test et qui ne doivent pas entrer dans la banque. L'utilisation du test auprès des apprenants suppose donc la mise en place d'un système d'administration.

4.1 L'unité de développement

4.1.1 *Données techniques*

Comme il nous semblait important d'élaborer un instrument informatisé qui soit non seulement intéressant du point de vue de la recherche mais qui puisse aussi servir éventuellement comme outil de classement auprès de la population visée, le choix d'un type d'appareils était déterminant. Nous avons éliminé dès le départ les gros systèmes fonctionnant en temps partagé. Bien que les premiers didacticiels aient été conçus pour de tels ordinateurs, la technologie des micro-ordinateurs personnels offre aujourd'hui beaucoup plus de possibilités pour les applications dans le domaine de l'éducation. Au moment où nous devions prendre la décision, quatre types d'appareils se partageaient le marché éducatif: les "Commodore", les "Apple" (série II), les "MacIntosh" et les IBM (ou leurs compatibles). Les deux premiers nous semblaient en perte de vitesse et l'arrivée du "MacIntosh" était trop récente pour qu'on puisse évaluer son impact

pour le marché de l'éducation. Nous avons donc opté pour la famille des IBM. La popularité qu'ont connu ces appareils par la suite nous a prouvé qu'il s'agissait là d'un choix judicieux.

Toutefois ce type d'appareils souffre parfois d'un manque de standardisation causé par la multiplicité des marques concurrentes et des configurations possibles. Afin d'assurer la compatibilité du logiciel, nous avons essayé de nous en tenir à un programme qui puisse fonctionner avec tous les appareils, y compris les modèles de base, sans requérir une installation particulière. Tout appareil IBM (ou compatible) doté d'une unité de disque (5¼ pouces) et de 256 K de mémoire vive, pourra exécuter le programme. Celui-ci fonctionne avec tous les types d'écrans et de cartes graphiques. L'utilisation d'un disque rigide permet toutefois d'accélérer l'exécution. Nous avons également produit une version, plus rapide et plus précise, pour les machines possédant un co-processeur numérique.

Essentiellement, le programme *CAPT* (*Computerized Adaptive Placement Test*) fonctionne comme un système de gestion d'une base de données auquel on a ajouté des fonctions spécifiques. Le programme est écrit en *Turbo-Pascal* (Borland 1987) et utilise *Turbo-Database*, un ensemble de sous-programmes mis au point et distribués par le fabricant du compilateur, pour la gestion des bases de données (Borland 1985b). Même si elle requiert une connaissance de la programmation et que le développement peut s'avérer fort coûteux, l'utilisation de ce type de logiciel présente des avantages considérables par rapport aux systèmes auteurs, trop fermés, ou par rapport aux logiciels habituels de bases de données (Henning 1986, pour un exemple avec *DBase*). D'une part, grâce à un langage de programmation, on peut intégrer les fonctions de gestion de la base de données à d'autres sous-programmes tels ceux reliés à l'administration du test. D'autre part, on obtient un programme compilé de sorte qu'il n'est pas nécessaire d'acheter, d'installer et de charger un système-auteur ou un interpréteur sur chacun des postes de travail.

Les fonctions de *Turbo-Database* permettent aussi la mise sur pied de banques considérables: plus de 65,000 enregistrements ou

items qui peuvent compter jusqu'à 64 K chacun. À l'instar de la plupart des logiciels de base de données, le système procède par indexation des enregistrements. De plus, des manupulations sophistiquées des structures de données arborescentes (technique *B+Tree*) permettent de retrouver les enregistrements désirés très rapidement. Dans le cas de l'application qui nous concerne, nous avons cherché un équilibre entre vitesse et volume en limitant chaque sous-test à une banque de 150 enregistrements (ou items). Chaque enregistrement occupe environ un K de mémoire.

La programmation s'est effectuée selon une approche modulaire. Cette approche respecte d'ailleurs les principes de programmation structurée que nous avons essayés de respecter tout au cours de la mise au point du logiciel. Nous croyons qu'il sera ainsi beaucoup plus aisé d'effectuer ultérieurement des modifications au logiciel. Le programme principal de l'unité de développement est donc relativement court. Il s'occupe simplement de faire apparaître l'écran de présentation, d'ouvrir les fichiers, d'amener les menus qui occupent une page-écran (menu principal et menu des opérations) et finalement de fermer les fichiers. Pour les autres opérations, le programme principal appelle une dizaine de fichiers inclus (les modules). Quatre d'entre eux proviennent directement de *Turbo-Database*:

- ACCESS.BOX: routines de base reliées aux fichiers de données et d'index;
- GETKEY.BOX: pour retrouver l'item désiré;
- ADDKEY.BOX: pour ajouter un item;
- DELKEY.BOX: pour éliminer un item.

Les autres modules ont été programmés pour les besoins spécifiques de l'unité de développement d'un test adaptatif:

- ÉCHELLE.INC: délimitation des niveaux;
- INOUT.INC: contrôle de l'affichage;
- INFO.INC: programme reliés à la fonction d'information;
- CONSULT.INC: consultation et mises à jour;
- EXTERN.INC: liste et transfert d'items;
- SIMUL.INC: simulation d'administration de tests adaptatifs.

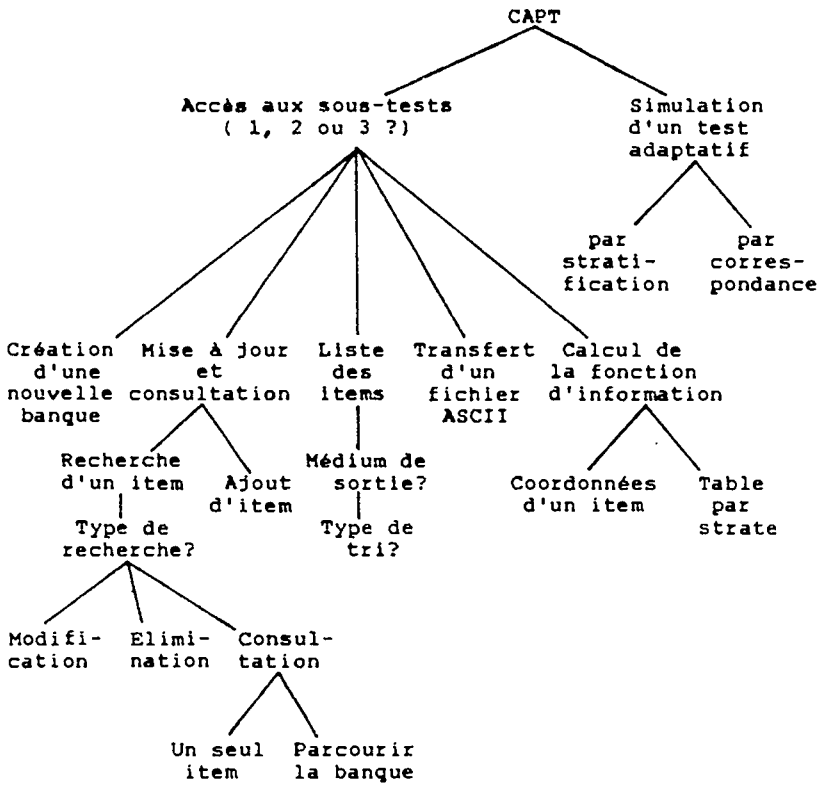
La programmation s'est faite à l'aide de la version 3.01 de *Turbo-Pascal*. Nous avons par la suite effectué une conversion vers la version 4.0 afin de mieux traiter les nombres très petits qui interviennent dans le calcul de l'habileté par maximum de vraisemblance. Nous avons toutefois conservé les fichiers inclus. Une fois compilée, l'unité de développement occupe environ 65 K de mémoire, sans compter la quinzaine de fichiers d'index et de données (5 par sous-test) auxquels ont accès les modules.

4.1.2 *Description des fonctions*

En concevant le système de développement, nous espérions mettre au point un logiciel qui jouerait trois rôles principaux. Premièrement, il devait, comme tout système de base de données, pouvoir entreposer un certain nombre de données concernant les items. On peut imaginer un grand nombre de renseignements succeptibles d'apparaître dans une banque d'items: formulation de la question, indices statistiques, numéro d'identification, date de création... On s'attend également à ce qu'on puisse accéder rapidement et facilement à ces renseignements et à ce qu'on puisse ajouter ou retrancher des items selon les besoins. Deuxièmement, il était important que le système ait la capacité de traiter une partie des données qui se trouvent dans la banque. Par exemple, dans la perspective de la théorie du trait latent, il nous semblait que lorsqu'on avait entré les paramètres obtenus lors de la calibration, le système devait être en mesure de fournir des renseignements sur l'information qu'apporte un item pour les points les plus pertinents de l'échelle d'habileté. Troisièmement, comme il s'agit en fin de compte d'en arriver à l'administration d'un test adaptatif, il était primordial que le système puisse simuler une séance de testing adaptatif et fournir des renseignements quant au déroulement.

En tenant compte de ces trois exigences, c'est-à-dire capacités de gestion de la banque, de traitement et de simulation, nous avons conçu un système dont les composantes s'organisent selon la hiérarchie de la figure 4.1.

FIGURE 4.1
Structure des fonctions de *CAPT*



Lorsqu'on appelle le programme *CAPT*, le menu principal apparaît à l'écran. L'usager choisit alors parmi les options suivantes:

- Accès au sous-test #1
- Accès au sous-test #2
- Accès au sous-test #3
- Simulation par stratification
- Simulation par correspondance
- Fin de la session

On voit donc que plutôt que d'entrer dans une des trois banques d'items, l'usager peut décider de simuler l'administration

d'un test adaptatif. Deux modes d'administration sont disponibles: par stratification ou par correspondance. Dans la section suivante, consacrée à l'unité d'administration, nous expliquerons plus en détail la différence entre ces deux modes d'administration. Au cours d'une simulation, l'usager joue le rôle d'un élève. Toutefois, il évite les consignes et les exemples et il peut suivre le déroulement du test. En effet, une fenêtre à l'écran lui permet de connaître, tout au cours du test, le nombre d'items réussis, l'estimation de son habileté par le programme, la quantité d'information accumulée... Il est également possible de faire imprimer les renseignements que contient la fenêtre en vue d'une analyse plus approfondie du déroulement du test.

Si l'usager choisit d'entrer dans une des trois banques d'items, le programme ouvre alors quatre fichiers: un fichier composé des enregistrements contenant les données sur les items, un autre comprenant la quantité d'information que donne chaque item pour 13 points de l'échelle d'habilité et deux fichiers d'indexation permettant de retrouver les items. On peut retrouver un item soit à partir du code d'identification que lui a assigné l'usager, soit à partir de l'indice *m* c'est-à-dire du point où l'item fournit le maximum d'information. L'indice *m* n'a pas à être entré par l'usager. Le programme se charge de le calculer à partir des valeurs que l'usager a attribué aux trois paramètres.

En entrant dans une des banque d'items, l'usager reçoit un autre menu, le menu des opérations, qui occupe lui aussi tout l'écran. Le menu des opérations énumère les possibilités d'intervention:

- Mise à jour de la banque
- Liste des items
- Transfert d'un autre fichier
- Table d'information
- Retour au menu principal

En choisissant la première possibilité, l'usager obtient tout d'abord une fiche où apparaissent les identificateurs des champs que comprend un enregistrement. On retrouve les champs suivants:

- le code d'identification de l'item
- la réponse correcte à la question
- l'incide *m* (calculé par le programme)
- le paramètre *a* (discrimination)
- le paramètre *b* (difficulté)
- le paramètre *c* (hasard)
- la formulation de la question (max. 6 lignes)
- les options de réponses (max. 6 lignes)

L'usager peut alors décider d'ajouter un item, c'est-à-dire de compléter la fiche, ou d'appeler un item qui se trouve dans la banque. La recherche d'item se fait soit par le code d'identification, soit par l'indice *m*. Dans ce dernier cas, le système affichera la fiche de l'item dont l'indice *m* se rapproche le plus de l'indice demandé. L'usager peut par la suite parcourir la banque en demandant l'item dont l'indice suit immédiatement (item plus difficile) ou précède (item plus facile) celui qui est affiché. Lorsqu'un item apparaît à l'écran, il est possible de le retirer de la banque ou d'en modifier un/plusieurs champ(s). La modification des champs, comme l'entrée des données, se font dans l'ordre de la fiche, la touche <Return> permettant de passer au champ suivant ou à la ligne suivante. On peut corriger en se servant des commandes habituelles d'effacement et de déplacement du curseur.

La seconde possibilité permet d'obtenir la liste (à l'écran, à l'imprimante ou sur disque) des items avec leur code d'identification, la réponse, l'indice *m*, les paramètres et le début de la question. L'usager peut décider d'ordonner cette liste selon le numéro de code ou selon l'indice *m*.

L'option «Transfert d'un autre fichier» permet d'importer plusieurs items rangés dans un fichier ASCII plutôt que de les entrer individuellement dans la banque. Il s'agit d'une caractéristique appréciable quand la version papier-crayon qui a servi à la calibration a été rédigée par traitement de texte. L'usager doit toutefois apporter certaines modifications au texte afin que le programme assigne correctement les segments de texte aux champs appropriés.

Enfin, il est possible de demander, pour un item particulier, les coordonnées de la courbe d'information pour le point central de chacun des sept niveaux que nous avons distingués de même que pour les six niveaux mitoyens. Nous avons en effet divisé la courbe normale de l'échelle d'habileté par strates comprenant théoriquement un nombre égal de sujets. Le point 0 correspond à la moyenne de la population c'est-à-dire au niveau intermédiaire moyen. On peut aussi obtenir un tableau montrant, pour chacune des strates, les dix items qui apportent le plus d'information. La table d'information est conservée dans un fichier qui doit être mis à jour si les paramètres des items changent.

Comme tout système de gestion de base de données, la structure d'un tel programe peut devenir vite très complexe. Toutefois, nous avons voulu en rendre l'utilisation la plus simple possible de manière à ce que, pour peu qu'il soit familier avec les concepts psychométriques, l'usager puisse s'y retrouver. De fait, il est possible d'emmagasiner des données pour tout type d'item à choix multiple, pré-calibré selon un modèle à trait latent. Dès que les trois banques sont en usage, on peut simuler l'administration d'un test adaptatif. D'autre part, comme on a respecté les principes de la programmation structurée, on pourrait facilement modifier le code-source pour l'adapter à une utilisation particulière[1].

4.2 L'unité d'administration

4.2.1 *Données techniques*

Alors que l'unité de développement sert à la mise sur pied de la banque, l'unité d'administration est le logiciel qu'on utilise avec les élèves. Tout comme l'unité de développement, l'unité d'administration fonctionne avec tous les ordinateurs personnels de la famille IBM. Conscient des problèmes que comporte l'emploi de ces appareils auprès de sujets qui ne s'en sont parfois jamais servis et pour qui ces appareils peuvent même inspirer de la crainte, nous

[1] Nous ne pouvons pas publier le code-source du programme. Toutefois de telles utilisations sont possibles en consultant l'auteur.

avons cherché à simplifier au maximum les interactions avec la machine. Ainsi grâce à un fichier auto-exécutable, le programme se charge automatiquement dès que la disquette est insérée et que l'appareil est mis en marche. En modifiant le fichier auto-exécutable, on peut adapter le programme selon des besoins ou des installations spécifiques: chargement à partir d'un disque rigide, intégration dans un réseau, utilisation à distance par modem... Sauf au moment où le sujet tape son nom, les seules touches qui déclenchent une action sont les choix de réponse (soit «a», «b», «c» et «d»), la barre d'espacement (pour omettre une réponse), la touche <Return> (pour commencer un sous-test) ou la touche <Escape> (pour interrompre l'exécution). Tout autre geste déclenche un signal sonore rappelant à l'élève de se limiter à ces touches.

Avec un appareil de type XT, les temps d'attente au début de chaque sous-test sont négligeables, voire imperceptibles; vers la fin du sous-test, quand l'appareil fait l'estimation du niveau par maximum de vraisemblance, ils deviennent sensiblement plus longs mais demeurent acceptables. Comme pour l'unité de développement, nous avons mis au point, parallèlement, une version pour des appareils dotés d'un co-processeur numérique. Puisque que ces appareils traitent plus efficacement les nombres très petits, le temps d'attente diminue considérablement et l'estimation est plus précise.

Par ailleurs, avec ou sans co-processeur numérique, le programme est disponible en deux modes d'administration différents: par stratification (programme STRAT) et par correspondance (programme MATCH). La différence tient à la façon dont le choix des items s'opère au cours de l'administration. Ainsi, même pour deux sujets d'habilité égale, il est possible que la sélection des items, ou tout au moins leur ordre, varie d'un mode à l'autre.

On peut voir l'unité d'administration comme un sous-ensemble de l'unité de développement. En effet, le programme emprunte en grande partie des portions du code du programme *CAPT* notamment à partir du module de simulation. Le programme principal se limite à l'écran de présentation et à l'enregistrement du

résultat. Les autres opérations sont prises en charge par des sous-programmes qui se trouvent dans des fichiers inclus (les modules). On a conservé deux des modules de *Turbo-Database*:

- ACCESS.BOX: routines de base reliées aux fichiers de données et d'index;
- GETKEY.BOX: pour retrouver l'item désiré;

On a programmé quatre modules:

- ÉCHELLE.INC: délimitation des niveaux;
- ÉCRAN.INC: contrôle de l'affichage;
- TUTOR.INC: exemples et directives à l'élève;
- ADMIN.INC: administration des composantes du test.

Une fois compilé, le programme compte un peu plus de 38 K. Dans la version actuelle, l'ensemble des fichiers nécessaires à l'administration occupe un peu moins de la moitié d'une disquette de 360 K. On pourrait donc doubler le nombre d'items et quant même disposer d'assez d'espace pour conserver les résultats d'environ 300 sujets.

4.2.2 *Algorithme d'administration*

Il convient pourtant de préciser certains aspects de l'algorithme d'administration.

4.2.2.1 *L'évaluation préliminaire*

Au début du test, on explique à l'élève ce qu'on attend de lui et la façon dont se déroule le test. On lui demande ensuite de taper son nom et son prénom puis de vérifier s'ils sont correctement inscrits. Le programme pose alors quelques questions à l'élève dans le but d'obtenir une première évaluation de son niveau à partir de ses contacts avec la communauté francophone du nombre d'années d'études du français et de sa propre évaluation.

On vérifie d'abord si le sujet a déjà vécu dans un milieu francophone. Si ce n'est pas le cas, le sujet n'accumulera pas de point à ce chapitre. Par contre, s'il a séjourné dans un milieu francophone entre 3 et 12 mois, il recevra 3 points alors que si la durée de son séjour dépasse 5 ans, il recevra le maximum soit 7.5. Le total est pondéré par la question suivante où l'on demande à l'étudiant d'indiquer la période qui s'est écoulée depuis qu'il a vécu dans ce milieu francophone: plus cette période est longue, plus le coefficient est faible. On demande ensuite à l'étudiant pendant combien d'années il a étudié le français. Chaque année d'étude au niveau secondaire compte pour un point et chaque année d'étude au niveau post-secondaire pour 1.5. Ici encore, le résultat est pondéré par le nombre d'années qui se sont écoulées depuis le dernier cours. La moyenne quant au nombre années de contact et de scolarité (avec un maximum de 6) forme près de la moitié de ce qui servira d'évaluation préliminaire. L'autre moitié vient de l'estimation que fait l'élève de son propre niveau. Oskarsson (1981) a démontré que des étudiants peuvent arriver à se situer assez bien par rapport à une échelle globale de niveaux en langue seconde. Le programme affiche les sept niveaux que nous reconnaissons et demande au sujet de déterminer lui-même son niveau de maîtrise générale du français. On assigne alors une valeur numérique à l'auto-évaluation (de 1 à 7) et on l'additionne aux points déjà amassés. Le nombre de points varie de 1 à 13, ce qui correspond au nombre de strates que nous distinguons.

Le score ainsi calculé sert de point de départ pour la sélection du premier item. Certaines études (Mussio 1973, Tung 1986) montrent que plus l'estimation de départ est précise, plus le choix des items sera approprié et plus l'estimation finale sera juste. Le premier sous-test utilise cette évaluation préliminaire; le second sous-test utilise le résultat du premier; le troisième fait la moyenne des deux premiers. Ce recours à des données déjà disponibles permet de réduire le nombre d'items à administrer pour atteindre le seuil d'information visé.

On peut certainement s'interroger sur le choix des critères retenus lors de l'évaluation préliminaire: contact, scolarité et auto-

évaluation. Cette décision a été guidée surtout par l'expérience auprés de la population visée et par le fait que ces données sont immédiatement disponibles et quantifiables. Au verso de la feuille de réponses de la version 2, on demandait aux sujets de fournir des renseignements sur leurs contacts avec le milieu francophone et sur leur études en français; on leur demandait aussi d'identifier le niveau auquel ils croyaient appartenir. En mettant en parallèle ces données avec les scores, on a déterminé l'importance relative de chaque critère. Cette façon de procéder par observation satisfaisait les besoins de la présente recherche. Il ne fait pas de doute néanmoins que ces données mériteraient d'être analysées plus systématiquement en vue de l'élaboration d'un modèle prédictif plus précis.

4.2.2.2 *La sélection des items*

De la même façon qu'il est avantageux de pouvoir choisir entre deux versions «papier-crayon» équivalentes, il peut être avantageux de pouvoir choisir entre deux versions adaptatives qui requièrent le même équipement, exploitent la même banque et demandent le même temps d'administration sans nécessairement appeler la même séquence d'items pour un niveau donné. Ainsi, selon qu'on opte pour l'administration par stratification ou l'administration par correspondance, le choix des items au cours du test pourra varier.

Le mode d'administration par stratification est une adaptation de la procédure mise de l'avant par l'équipe de Weiss (Vale et Weiss 1975) et présentée au début du second chapitre. Lorsqu'un item est entré dans la banque à l'aide de l'unité de développement, on range, dans un fichier, la quantité d'information qu'apporte cet item pour les treize points de l'échelle d'habileté qui nous intéressent. Ces données servent par la suite à constituer une table d'information, c'est-à-dire une matrice de 13 x 10 où pour chacune des treize «strates» ainsi déterminées, on range le code d'identification des dix items qui apportent le plus d'information à ce niveau. Les items sont ordonnés selon l'ordre décroissant de la

quantité d'information qu'ils apportent, l'indice le plus petit étant attribué à l'item le plus efficace. Du point de vue de la structure des données au plan informatique, chaque strate peut être considérée comme une queue. Lors de l'administration, on détermine la strate à partir de l'estimation provisoire du niveau d'habileté du sujet et on vérifie si l'item suivant a été utilisé; si c'est le cas on poursuit la recherche, sinon on l'appelle avec son code puis on l'identifie comme item déjà utilisé. La procédure se poursuit jusqu'à ce qu'une des quatre conditions suivantes soit remplie:

- la queue est vide;
- le sujet a obtenu un score parfait (10/10);
- le sujet a obtenu un score nul (0/10);
- on atteint le degré de précision requis.

Le degré de précision correspond, comme on le sait, à l'erreur type maximale permise laquelle est une fonction inverse de la quantité d'information amassée. Pour l'administration par stratification, ce seuil a été fixé à .35. Cette marge d'erreur est légèrement inférieure à la moyenne de l'erreur type qu'on trouve pour différents niveaux avec la version papier-crayon. Normalement, le sous-test se termine lorsqu'on a atteint le degré de précision requis. En fixant le seuil à .35, on s'assure d'une fiabilité générale au moins équivalente à celle de la version «papier-crayon». Par ailleurs, compte tenu du nombre d'items dans la banque et de leur discrimination, il est difficile de minimiser davantage l'erreur. Il faut noter que la première des quatre conditions est rarement remplie car le nombre d'items dans la queue a été déterminé en tenant compte de l'erreur type maximale et de la taille de la banque d'items.

Comme nous l'avons mentionné au deuxième chapitre, le concept de l'administration par correspondance vient de Lord (1970, 1977). Birnbaum (1968) avait déjà établi que, pour un modèle à trois paramètres, on pouvait calculer une valeur m_i sur le continuum de l'échelle d'habileté, correspondant au maximum de la fonction d'information pour un item i. On utilise alors la formule 4.1:

$$m_i = b_i + \frac{1}{Da_i} \ln\left[\frac{1 + \sqrt{1 + 8c_i}}{2}\right] \qquad (4.1)$$

L'indice *m* représente donc le point de la courbe d'information où l'item est le plus efficace. Quand on entre les paramètres d'un item avec l'unité de développement, le programme calcule l'indice *m* et l'inscrit dans un fichier d'index. Au moment de l'administration par correspondance, il suffit donc de chercher l'index qui se rapproche le plus de l'estimation provisoire de l'habileté du sujet. On peut alors appeler l'item approprié en utilisant cet index. On répète l'opération aussi longtemps qu'on n'a pas compté dix réponses toutes incorrectes ou toutes correctes ou, comme c'est généralement le cas, aussi longtemps qu'on n'a pas atteint le degré de précision requis. Il faut souligner qu'afin de s'assurer que la durée du test par correspondance soit comparable à celle du test par stratification nous avons dû hausser le maximum de l'erreur type pour le porter à .5. Avec un nombre égal d'items, le test par correspondance est donc moins fiable. Cela tient au fait que la procédure utilise une plus grande variété d'items y compris certains dont le paramètre de discrimination peut être assez bas. La procédure par stratification, au contraire, tend à ne retenir dans la table que les items qui sont les plus discriminants. Nous reviendrons sur les implications du choix du mode d'administration dans le dernier chapitre.

4.2.2.3 *L'estimation de l'habileté*

Théoriquement, l'adéquation d'un ensemble d'items à un modèle à trait latent implique qu'on puisse soumettre des items différents à plusieurs sujets et pouvoir tout de même comparer leurs résultats. Cela suppose qu'on dispose d'une procédure permettant de déterminer à quel degré d'habilité une configuration de réponses donnée est la plus susceptible de se rencontrer. En d'autres termes, il faut trouver la valeur de Θ_i qui maximise la fonction de vraisemblance. Cette fonction de vraisemblance L où U_i prend la valeur 0 (réponse incorrecte) ou 1 (réponse correcte) peut s'exprimer ainsi:

$$L(U_1, U_2 \ldots U_n) = \prod_{i=1}^{n} P_i^{U_i} \; Q_i^{1-U_i} \qquad (4.2)$$

Il est plus commode cependant d'utiliser le logarithme naturel de la fonction qui peut ainsi se réécrire en 4.3:

$$\ln L(U|\theta) = \sum_{i=1}^{n} [U_i \ln P_i + (1 - U_i) \ln Q_i] \qquad (4.3)$$

On atteint le maximum de vraisemblance lorsque Θ permet d'assigner la valeur 0 à la première dérivée:[2]

$$\frac{d}{d\theta} \ln L(U|\theta) = 0 \qquad (4.4)$$

Pour résoudre cette équation, on doit recourrir à des procédures numériques comme la méthode Newton-Raphson. Essentiellement, la méthode consiste à soustraire un facteur de correction (le rapport entre la première et la seconde dérivée de la fonction de vraisemblance) à une première estimation de l'habilité. On reprend la procédure de façon itérative sur la nouvelle estimation ainsi obtenue jusqu'à ce que le facteur de correction soit négligeable. On dit alors qu'il y a convergence.

Cette procédure n'est pas sans problème. Au plan théorique, le problème le plus sérieux tient au fait que la solution n'est pas toujours unique, surtout avec un nombre réduit d'items. Ainsi le programme doit rejeter toute solution qui assignerait à Θ une valeur hors de la gamme des valeurs que cette variable est susceptible de prendre. De plus, la procédure ne peut pas s'appliquer lorsqu'on se trouve devant un score nul ou un score parfait. C'est pourquoi on doit attendre d'avoir obtenu cinq réponses, dont au moins une correcte et une incorrecte, avant de procéder à l'estimation par niveau de vraisemblance. Tant qu'on ne peut pas calcuier le maximum de vraisemblance, l'estimation du niveau s'effectue d'une façon mécanique. Avec l'administration par stratification, on passe à la strate supérieure si la réponse est exacte ou à la strate inférieure dans le cas contraire. Avec l'administration par correspondance, on augmente ou diminue Θ d'environ .2. Avec les deux modes d'administration, on essaie de tenir compte de l'effet du hasard lorsque le sujet fournit une réponse exacte. Par ailleurs, après dix questions, le programme classe comme «Débutant» (niveau 01), le sujet qui n'en a réussi aucune et comme «Très avancé +» (niveau 14) celui qui les a toutes réussies.

---

[2] Hambleton et Swaminathan décrivent les étapes dans l'estimation de l'habileté et donnent la formule à employer pour obtenir les deux premières dérivées.

Il faut également noter que l'utilisation d'un micro-ordinateur pose des problèmes particuliers. D'une part, l'erreur d'arrondissement peut faire dériver l'itération et ainsi être responsable du fait n'y ait pas de convergence. D'autre part, la procédure demande tellement de temps qu'il nous a fallu réduire le nombre d'itérations à 25 pour les estimations provisoires et à 50 pour l'estimation finale. Si après 50 itérations, il n'y a toujours pas de convergence ou que le résultat est hors des limites prévues, on a recours à une procédure alternative qui consiste simplement à considérer la moyenne du niveau (la strate) des items réussis par rapport à la moyenne du niveau des items échoués. La moyenne de ces deux valeurs correspond à l'estimation de l'habilité du sujet. Cette technique se rapproche de la technique *up and down* dont discute Mussio (1973) dans une étude sur la sélection des items.

Certes, on pourrait rendre plus complexe et plus précise la méthode pour estimer le niveau des sujets. On pourrait, entre autre, intégrer une évaluation bayésienne du type de celle que décrit Owen (1975), faire appel à des procédures mieux adaptées aux micro-ordinateurs (Bock et Mislevy 1982), ou chercher à obtenir des estimations provisoires plus précises (Reckase 1983). Toutefois, puisque dans sa version actuelle, le test fonctionne de façon tout à fait acceptable, de tels raffinements dépassent sans doute les objectifs de notre étude.

4.2.2.4 *Le résultat final*

Avec la procédure pour l'estimation de l'habilité que nous venons de décrire, le résultat final est obtenu, soit en appliquant la technique du maximum de vraisemblance (avec un maximum de 50 itérations), soit en considérant les moyennes de niveaux pour les items réussis et pour les items échoués. Nous avons observé qu'une fois sur quatre, il n'y a pas de convergence lors de l'évaluation finale et que la convergence semble plus difficile à obtenir lors de l'administration par stratification.

Toutefois la technique alternative pour l'estimation finale nous semble particulièrement approprié dans la mesure où d'une part, elle tient compte des estimations provisoires où la convergence s'est réalisée et d'autre part, elle exprime le résultat en terme de niveau. En effet, c'est le résultat de l'étudiant par rapport aux quatorze niveaux distingués qui nous intéresse plutôt que l'estimation de l'habilité par rapport à la courbe normale. Voilà pourquoi, le résultat final est défini comme la moyenne des niveaux atteints aux trois sous-tests. Compte tenu qu'il n'y avait que trois sous-tests, il nous semblait superflu de chercher à établir, comme le suggère Weiss et Brown (1978), une régression multiple entre les sous-tests afin de déterminer le niveau. Ainsi, le sujet qui a été classé «Faux débutant +» (04) au premier sous-test, «Intermédiaire II» (07) au deuxième sous-test et «Intermédiaire I» (05) au troisième sous-test, sera finalement classé au niveau «Intermédiaire I (05). Comme pour la version «papier-crayon», on donne ainsi la même pondération au trois sous-tests.

Le test se termine par un message de remerciement qui accompagne le résultat final. Ici encore, on communique le niveau plutôt qu'une note ou un indice statistique qui aurait peu de signification pour l'étudiant. Ce résultat est rangé dans un fichier qui pourra par la suite être réordonné et imprimé en le remaniant à l'aide d'un logiciel de traitement de texte ou en utilisant simplement les commandes standard du système d'exploitation (SORT et PRINT). Il est alors possible d'obtenir rapidement une liste des sujets qui ont fait le test, soit par ordre alphabétique, soit par niveau. On peut penser qu'une telle liste devrait faciliter grandement la création des groupes-classes.

4.3 La mise à l'essai

4.3.1 *Exemple du déroulement d'un test*

Christine T. étudie à l'université et y a suivi deux cours de français langue seconde. Comme elle n'a jamais vécu en milieu francophone, elle ne se croit pas capable de fonctionner en français. Installée devant un micro-ordinateur qui exécute le programme STRAT, Christine a donné ces renseignements à la machine. Celle-ci,

pour l'instant juge la candidate «Faux débutant +» et sélectionne en conséquence le premier item du premier sous-test (compréhension). Il s'agit de l'item CO36, comme le montre le tableau 4.1 qui reproduit la ligne d'information que peut imprimer le programme de simulation. Chistine T. tape la bonne réponse de sorte que le programme lui présente un item plus difficile, le CO24. La procédure se poursuit jusqu'au cinquième item, le CO45. À ce moment, le programme fait l'estimation du niveau d'habileté de Christine et interrompt ce sous-test car le seuil d'information visé est déjà atteint. Après ce premier sous-test, Christine se situe au niveau «Intermédiaire I» (Θ = -.216). C'est à ce niveau, que correspondra donc l'item EA34, le premier du deuxième sous-test (énoncé approprié).

TABLEAU 4.1
Simulation par stratification

Test	Item Courant	Dernier item	Réussis	Total	Theta	Info.	Erreur	Itér.
# 1	CO36		0	0	-0.500	?	?	0
# 1	CO24	CO36	1	1	-0.300	?	?	0
# 1	CO25	CO24	2	2	-0.300	?	?	0
# 1	CO39	CO25	2	3	-0.100	?	?	0
# 1	CO50	CO39	2	4	-0.300	?	?	0
# 1	CO45	CO50	3	5	-0.035	?	?	14
# 1		CO45	4	6	-0.216	8.582	0.341	7
# 2	EA34		0	0	-0.216	?	?	0
# 2	EA20	EA34	1	1	-0.300	?	?	0
# 2	EA19	EA20	2	2	-0.100	?	?	0
# 2	EA05	EA19	2	3	-0.300	?	?	0
# 2	EA10	EA05	2	4	-0.500	?	?	0
# 2	EA17	EA10	3	5	-0.175	?	?	10
# 2	EA15	EA17	3	6	-0.178	?	?	3
# 2	EA43	EA15	3	7	-0.413	3.885	0.507	14
# 2	EA30	EA43	3	8	-0.212	4.527	0.470	25
# 2	EA42	EA30	3	9	-0.300	4.826	0.455	18
# 2	EA36	EA42	4	10	-0.548	4.714	0.461	17
# 2	EA40	EA36	5	11	-0.413	4.971	0.449	19
# 2		EA40	5	12	-0.393	5.508	0.426	13
# 3	CL06		0	0	-0.305	?	?	0
# 3	CL26	CL06	1	1	-0.100	?	?	0
# 3	CL32	CL26	1	2	-0.300	?	?	0
# 3	CL28	CL32	2	3	-0.100	?	?	0
# 3	CL22	CL28	3	4	-0.100	?	?	0
# 3	CL50	CL22	4	5	0.249	?	?	12
# 3	CL44	CL50	5	6	0.055	7.481	0.366	9
# 3	CL46	CL44	6	7	0.009	8.020	0.353	7
# 3	CL30	CL46	7	8	0.613	7.035	0.377	21
# 3		CL30	8	9	0.303	11.825	0.291	17

Niveau général: Intermédiaire I +

Cette fois, comme les paramètres de discrimination sont plus bas, le programme présentera le nombre maximum d'items, soit 12, sans pour autant atteindre le seuil d'information visé. À la fin du deuxième sous-test, l'estimation de l'habileté n'a pas changé sensiblement ($\Theta = -.393$) et c'est donc au niveau «Intermédiaire I» que commence le troisième sous-test (phrases à trou), avec l'item CL06. Comme Christine réussit bien dans ce genre d'exercice, les items deviennent rapidement plus difficiles. Après 9 items, le programme estime que pour cette partie, la candidate se situe au niveau «Intermédiaire III». En considérant l'ensemble des trois sous-tests, la machine informe Christine qu'elle a été classée au niveau «Intermédiaire I +» (à la frontière entre l'intermédiaire faible et le moyen).

TABLEAU 4.2
Simulation par correspondance

Test	Item Courant	Dernier item	Réussis	Total	Theta	Info.	Erreur	Itér.
# 1	CO50		0	0	-0.500	?	?	0
# 1	CO33	CO50	1	1	-0.320	?	?	0
# 1	CO39	CO33	2	2	-0.150	?	?	0
# 1	CO24	CO39	2	3	-0.375	?	?	0
# 1	CO36	CO24	3	4	-0.200	?	?	0
# 1	CO37	CO36	4	5	-0.025	?	?	Max.
# 1	CO34	CO37	5	6	0.150	?	?	Max.
# 1		CO34	5	7	-0.207	8.055	0.352	11
# 2	EA19		0	0	-0.207	?	?	0
# 2	EA49	EA19	0	1	-0.432	?	?	0
# 2	EA45	EA49	0	2	-0.657	?	?	0
# 2	EA36	EA45	0	3	-0.882	?	?	0
# 2	EA05	EA36	1	4	-0.707	?	?	0
# 2	EA10	EA05	1	5	-0.932	?	?	Max.
# 2	EA34	EA10	2	6	-0.757	?	?	Max.
# 2	EA13	EA34	3	7	-0.582	?	?	Max.
# 2	EA48	EA13	4	8	-0.407	?	?	Max.
# 2	EA43	EA48	5	9	-0.232	?	?	Max.
# 2	EA32	EA43	5	10	-0.457	?	?	Max.
# 2	EA03	EA32	6	11	-0.282	?	?	Max.
# 2		EA03	7	12	-0.500	?	?	Max.
# 3	CL43		0	0	-0.354	?	?	0
# 3	CL15	CL43	1	1	-0.179	?	?	0
# 3	CL06	CL15	2	2	-0.004	?	?	0
# 3	CL13	CL06	3	3	0.171	?	?	0
# 3	CL18	CL13	4	4	0.346	?	?	0
# 3	CL36	CL18	4	5	1.197	?	?	22
# 3	CL11	CL36	4	6	0.408	2.493	0.633	20
# 3	CL02	CL11	5	7	0.907	2.025	0.703	24
# 3	CL38	CL02	5	8	0.592	3.008	0.577	18
# 3	CL04	CL38	5	9	0.712	3.977	0.508	20
# 3		CL04	6	10	-0.164	4.100	0.494	19

Niveau général: Intermédiaire I

Le tableau 4.1 rend compte du déroulement d'une séance de testing adaptatif selon la procédure par stratification (STRAT). Au tableau 4.2, nous reproduisons ce qu'on obtient en simulant un test par correspondance (MATCH). Dans ce cas-ci, Christine serait plutôt classé «Intermédiaire I». On remarque qu'avec cet algorithme, on obtient moins d'information. Il faut aussi noter que cette fois, au deuxième sous-test, l'estimation par maximum de vraisemblance a échoué puisqu'après 50 itérations (le maximum), il n'y a toujours pas convergence. On doit donc recourir à une procédure alternative et ainsi déterminer le niveau pour ce sous-test d'après le rapport entre le niveau moyen des items réussis et le niveau moyen des items manqués.

4.3.2 *Originalité du système*

Le French CAPT est un didacticiel de testing adaptatif original et opérationnel. C'est à notre connaissance le seul test adaptatif qui ait pour but d'évaluer la maîtrise générale du français comme langue seconde. Sa banque a été mise sur pied non pas à partir de réponses simulées mais de sujets réels. Si le test est utilisé de façon régulière dans un établissement, on pourra amasser des données intéressantes sur la pertinence du test et sur ses effets auprès des étudiants. En ce sens, la comparaison que nous faisons dans le chapitre suivant n'est que l'amorce d'une série de recherches qui restent à faire sur l'utilisation d'un test adaptatif en langue seconde.

Le test se distingue de plusieurs de ses prédécesseurs par certains raffinements. D'abord, contrairement à plusieurs tests adaptatifs expérimentaux, le test que nous avons mis au point comprend plusieurs sous-tests. Ensuite, nous avons conçu un système basé sur une modèle à trois paramètres plutôt que sur le modèle de Rasch. Ce dernier, bien que beaucoup plus commode, nous semblait en effet mal convenir au type de test qu'on peut imaginer faire avec un ordinateur. Enfin, nous offrons deux procédures de sélection d'items (par stratification et par correspondance) qui devraient permettre de choisir des items différents pour des sujets de même niveau.

4.3.3 *Aspects à améliorer*

Comme pour tous les logiciels, il faut s'attendre à ce que de nouvelles versions voient le jour. Nous avons déjà mentionné quelques problèmes au plan de l'estimation du niveau d'habileté. Il est certain que nous devrons nous pencher sur cette question et programmer une procédure plus efficace qui tienne mieux compte des derniers développements en psychométrie et des limites, sans cesse repoussées mais toujours présentes, du micro-ordinateur.

Par ailleurs, le manque de temps, d'équipement ou de connaissances techniques nous a parfois conduit à certaines simplifications qui pourraient être corrigées. De plus, lors de la mise à l'essai, nous avons pu observer quelques déficiences auxquelles il faudra éventuellement remédier. Nos priorités en ce qui a trait aux versions futures sont, par ordre d'importance, les suivantes:

- Bloquer temporairement l'entrée des réponses: Certains étudiants, surtout avec des claviers très sensibles, maintiennent le doigt trop longtemps sur les touches et provoquent ainsi un emballement du programme. Il s'agit d'une correction relativement simple à apporter en utilisant l'horloge interne de l'appareil.
- Rendre la touche <Shift> inopérante: La plus grande difficulté que rencontrent les étudiants dans l'interaction avec la machine se trouve au moment où ils tapent leur nom. En inscrivant automatiquement celui-ci en majuscules, on devrait simplifier la tâche à l'élève. Ici encore, il s'agit d'une modification mineure.
- Améliorer la présentation du résultat final: Au lieu de simplement afficher l'étiquette correspondant au niveau final, il serait souhaitable d'améliorer la présentation visuelle du résultat; on pourrait représenter celui-ci sous la forme d'un «baromètre» de sorte que l'étudiant puisse se situer par rapport à l'ensemble de la population.
- Recalibrer avec des sujets supplémentaires: Nous conservons les feuilles de réponses aux versions «papier-crayon»

3.1 et 3.2 qui sont présentement en usage. Nous espérons de la sorte recalibrer la banque d'items avec au moins un millier de réponses par item.

- Élargir la banque: Les trois banques comptent présentement un total de 127 items. Nous estimons qu'on pourrait grandement améliorer la qualité de la mesure sans compromettre l'efficacité de l'administration en ayant de 200 à 250 items. Pour calibrer des items supplémentaires, il nous faudra créer des versions «papier-crayon» avec des items d'ancrage choisis parmi les meilleurs items déjà calibrés. Ce processus pourrait aussi servir à épurer la banque de certains items moins satisfaisants, particulièrement dans le deuxième sous-test.
- Rendre la présentation plus attrayante: La présentation du test est pour l'instant assez terne. En exploitant la couleur de même que les possibilités graphiques et musicales de l'appareil, on rendrait le test plus agréable. Toutefois, cette opération pourrait nous obliger à mettre au point plusieurs versions afin de tenir compte des diverses configurations d'équipement.
- Limiter les cas frontières: Pour l'instant, la courbe normale est divisée en strates égales. On pourrait toutefois envisager de réduire le nombre de sujets classés à des niveaux mitoyens (ex: «Avancé +») en retrécissant la bande et en diminuant le niveau de l'erreur type acceptable pour les sujets qui se trouvent classés à ces niveaux. Ainsi, ceux qui y resteraient seraient de véritables cas frontières c'est-à-dire des sujets qui pourraient aussi bien être classés au niveau supérieur qu'au niveau inférieur.
- Réviser la formule de l'évaluation préliminaire: En analysant statistiquement les données recueillies quant à l'expérience antérieure en français, on pourrait découvrir la contribution réelle des trois facteurs que sont le contact, la scolarité et l'auto-évaluation. On pourrait alors mettre au point une formule plus juste.

— Ajouter un sous-test de compréhension: Pour le classement dans un cours où la composante orale est importante, on s'attendrait à disposer, sinon d'un sous-test d'expression orale, du moins d'un test de compréhension auditive. Les observations que nous avons pu faire en vérifiant la validité de notre instrument nous portent à penser qu'un tel sous-test pourrait servir à évaluer la maîtrise générale. Sans recourir à des technologies aussi lourdes et coûteuses que le vidéodisque, on pourrait exploiter la possibilité de coupler le micro-ordinateur à un magnétophone ou à des supports de son numérisé (ex.: *CD-ROM*).

On voit donc que l'instrument dans sa version actuelle n'est pas parfait et on peut penser que chaque amélioration ouvrira la voie à des raffinements qu'on n'avait pas soupçonnés. Cependant, nous estimons que le test est, dans sa version actuelle, tout à fait utilisable et qu'avant de l'améliorer, il faut évaluer si l'investissement de temps, d'argent et d'énergie que cela suppose, est justifié.

⑤

LA COMPARAISON D'UN POINT DE VUE THÉORIQUE

Le projet de mettre sur pied un test adaptatif s'appuie sur l'hypothèse que cette formule permet, dans certaines conditions, une évaluation plus juste et plus commode de la performance de l'apprenant en langue seconde. Notre démarche a consisté à d'abord élaborer un test «papier-crayon» pour en faire par la suite une version informatisée. Reste la question fondamentale: quels sont les avantages du passage d'une version à l'autre?

Nous tenterons d'abord de répondre à cette question d'un point de vue théorique en faisant l'inventaire de ce qu'on peut imaginer comme avantages et inconvénients du testing adaptatif aussi bien au plan psychométrique, qu'au plan psychologique, qu'au plan administratif. Nous nous inspirons des exposés de Larson et Madsen (1985) et de Tung (1986) pour ce chapitre qui se veut le fruit d'une réflexion sur la question plutôt que le résultat d'une expérimentation.

5.1 Les avantages du testing adaptatif

5.1.1 *Au plan psychométrique*

Ce sont essentiellement des considérations d'ordre psychométrique qui ont inspiré les recherches autour du concept de «testing adaptatif». Théoriquement, le testing adaptatif permettrait en effet d'améliorer la qualité de la mesure en éliminant certaines sources de variance indésirables.

— **Fiabilité étendue**: Weiss (1982) souligne que beaucoup de ceux qui élaborent des tests conventionnels sont confrontés au dilemme «largeur de bande vs fiabilité». En effet, si on désire obtenir une mesure suffisamment précise à un niveau particulier sans prolonger indûment la séance, il faut renoncer à des items qui mesurent à d'autres niveaux. Il en résulte que même dans des tests qui s'adressent à une population où les niveaux varient beaucoup, comme dans le cas d'un test de classement, les tests conventionnels négligent généralement les points extrêmes de l'échelle d'habileté. Théoriquement, avec le testing adaptatif, en maintenant la même marge d'erreur acceptable, on obtient une distribution rectangulaire de l'information de sorte que le test est aussi précis à un niveau qu'à un autre.

— **Fiabilité accrue**: La fiabilité du testing adaptatif a fait l'objet de plusieurs études dont nous rapporterons les conclusions plus loin. Comme la procédure de testing adaptatif sélectionne les items les plus pertinents, il en résulte que pour un nombre égal d'items, la mesure à un niveau particulier sera plus précise que celle obtenue avec un test conventionnel commun à tous les candidats. On peut même imaginer des systèmes où l'on ferait varier la marge d'erreur acceptable selon le niveau. Par exemple, dans un test de classement, on peut choisir d'augmenter la fiabilité avec les cas-frontières; dans un test de certification, on peut concentrer l'information autour du seuil de passage.

— **Mode de correction plus efficace**: À moins de disposer de l'équipement nécessaire pour effectuer une correction électronique, on corrige habituellement les tests conventionnels en comptant le nombre de réponses correctes. Ce mode de correction n'est valable que si tous les sujets répondent aux mêmes items et qu'il est raisonnable de croire que chaque item doive recevoir la même pondération. Par contre, dans le cadre d'un test adaptatif, une fois les paramètres des items bien identifiés, on peut effectuer une correction par maximum de vraisemblance. Cette technique doit théoriquement donner la meilleure estimation de l'habileté d'un sujet. Certains problèmes inhérents au calcul par maximum de vraisemblance en limitent parfois l'utilisation: double solution, non

convergence, score nul ou parfait... Dans ces cas, on peut recourir à des solutions alternatives (moyenne des strates ou correction bayésienne) qui tiennent aussi mieux compte des configurations de réponses et des paramètres de chaque item que le mode de correction conventionnel.

— **Validité supérieure**: Comme la procédure de testing adaptatif conduit à soumettre à l'étudiant des questions qui correspondent davantage à ce qu'il peut ou sait faire, on peut croire que les tâches auxquelles il sera confronté seront plus réalistes. De ce point de vue, on améliore la validité de l'instrument car on s'assure de mesurer ce qui peut être effectivement mesuré à un niveau d'habileté donné.

— **Nouveaux types d'items**: L'ordinateur offre des ressources dont on ne dispose pas avec les questionnaires traditionnels notamment en ce qui a trait aux possibilités graphiques de l'écran. On peut donc imaginer de nouveaux types d'items qui pourraient correspondre davantage à l'objectif poursuivi. Dans l'élaboration de tests psychologiques, Cory et Rimbald (1977) ont ainsi remarqué que l'ordinateur permettait de convevoir des items originaux qui mesuraient mieux la mémoire à court terme et le raisonnement séquentiel. Il faut noter que dans le test *CAPT*, tous les items ont été calibrés à partir d'une version «papier-crayon» et que cette approche ne convient plus si on désire innover en construisant des items qui utilisent des ressources propres à la machine.

— **Détection de configurations de réponses inhabituelles**: Au cours de l'administration d'un test adaptatif, il est possible d'intégrer des procédures de calcul d'indices d'adéquation tels que ceux que proposent Tatsuoka et Tatsuoka (1982) ou Levine et Drasgow (1983). Nous ne voyions pas le besoin d'intégrer de telles procédures dans le test *CAPT*. Cependant, il est certain que la détection des réponses aberrantes est importante lorsqu'on a des raisons de douter de l'honnêteté ou du sérieux certains candidats ou qu'on constate des différences socio-culturelles marquées à l'intérieur de la population.

— **Traitement des différences socio-culturelles**: La détection des réponses aberrantes ne sert qu'à repérer les sujets pour qui le test ne convient pas. Toutefois, si un sujet est identifié au début d'une session de testing adaptatif comme membre d'un groupe minoritaire pour qui le test pourrait être biaisé, on peut tenir compte de cet aspect lors de la sélection des items. Pine et Weiss (1978) démontrent l'efficacité de la «prédiction différentielle» pour administrer des tests plus justes à une population minoritaire de race noire. La technique consiste essentiellement à procéder à des calibrations indépendantes pour chaque groupe et à utiliser le jeu de paramètres le plus approprié. Il faut souligner que compte tenu de l'homogénéité de notre population, et de la complexité de la programmation, nous n'avons pas recouru à ce raffinement supplémentaire.

— **Traitement de l'information préalable**: L'appartenance à un sous-groupe est l'une des données que l'on peut considérer au départ. De fait, une procédure de testing adaptatif efficace peut aussi tenir compte d'une variété de données telles que l'exposition à la langue seconde, le dossier scolaire, l'auto-évaluation. Il peut aussi s'agir des résultats des sous-tests antérieurs dans le cas de tests à plusieurs sections. Toute cette information peut être considérée au début d'un test afin de trouver, dès le départ, un item pertinent.

5.1.2 *Au plan psychologique*

Dès qu'elle fut lancée, l'idée du testing adaptatif en a séduit plusieurs du fait qu'elle faisait miroiter la perspective de varier le test selon l'apprenant, de viser «une mesure sur mesure». La plupart des avantages qu'on peut voir dans le testing adaptatif sont reliés à cette possibilité d'individualiser l'administration.

— **Élimination des items trop difficiles**: On connaît le sentiment de frustration que peut vivre l'étudiant débutant qui doit subir une série d'items beaucoup trop difficiles. Cette frustration devient vite du découragement, l'étudiant voit l'apprentissage d'une langue comme un objectif inatteignable et se culpabilise même de son ignorance. Il est fréquent qu'il ne complète pas le test. Le

sentiment de frustration peut être partagé par tout sujet à qui on présente des items trop difficiles. Il faut toutefois souligner que la norme prescrite par la théorie (50% de chance de réussite quand le hasard ne joue pas) ne correspond pas nécessairement au seuil psychologique auquel se réfère un étudiant pour juger de la difficulté d'un item. Prestwood et Weiss (1977) observent en effet que les étudiants faibles jugent souvent les items de leur niveau trop difficiles.

— **Élimination des items trop faciles**: Inversement, l'étudiant avancé à qui on soumet des items trop faciles aura l'impression de perdre son temps et jugera que le test ne lui rend pas justice. L'absence de défi se traduit par une perte d'intérêt et éventuellement par des réponses erronées parce que le sujet n'arrive plus à se concentrer sur une tâche qu'il estime de toute façon trop facile voire futile.

— **Correction immédiate**: On peut programmer le système de sorte que l'étudiant sache s'il a répondu correctement sitôt sa réponse tapée. Il est permis de croire que les sujets apprécient de savoir s'ils ont bien répondu et que cette rétroaction instantanée peut jouer un rôle important dans l'optique d'une évaluation formative. Il n'est pas certain toutefois que cette rétroaction instantanée ait toujours des effets positifs d'autant plus qu'elle peut affecter l'indépendance des items et mener à la divulgation des réponses. Par contre, sans nécessairement révéler la réponse des chaque question, on peut communiquer à l'étudiant son résultat final dès que le test est terminé. L'étudiant n'a donc pas à faire de démarches supplémentaires pour obtenir son résultat ou à attendre que les correcteurs aient achevé leur travail. Avec le système *CAPT*, on affiche simplement le niveau auquel l'étudiant a été classé, mais il est clair que ce message pourrait être plus nuancé. On pourrait, par exemple, indiquer le résultat pour chaque sous-test ou indiquer à l'élève le(s) cours qu'il pourrait suivre.

— **Test personnalisé**: Que ce soit parce que la machine interpelle le sujet par son nom, ou parce que le message varie d'une

situation à l'autre, ou encore parce que le programme semble tenir compte des réponses du sujet, celui-ci aura l'impression que le test est fait pour lui. Il appréciera qu'on ne lui impose pas l'anonymat des tests conventionnels.

— **Environnement facilitant**: Dès 1973, Johnston et Mihal avaient remarqué que l'administration d'un test par ordinateur permettait aux membres des minorités noires de mieux réussir. Saracho (1987) signale aussi que l'enseignement assisté par ordinateur peut favoriser l'apprentissage chez les élèves de groupes minoritaires. De fait, le contexte de l'administration d'un test informatisé se distingue de celui d'un test conventionnel traditionnellement associé aux valeurs qu'impose la majorité dominante dans le système d'éducation.

— **Administration sur demande**: Bien que relié davantage à l'utilisation du test adaptatif qu'à ses propriétés intrinsèques, le fait que l'étudiant ait l'occasion de faire le test quand il en a envie ou quand il se sent prêt à le faire, peut, dans certaines circonstances, rendre la formule du test adaptatif particulièrement attrayante pour l'étudiant.

— **Aspect ludique**: Il ne s'agit pas non plus d'une propriété intrinsèque au test mais il est certain que l'engouement que connaît l'ordinateur à des fins de divertissement peut contribuer à faire percevoir le test comme un jeu plutôt qu'une épreuve. Il peut en effet être amusant d'interagir avec une machine.

5.1.3 *Au plan administratif*

Au plan administratif, plusieurs des avantages qu'on peut imaginer s'apparentent aux considérations qui entrent en jeu lorsqu'un organisme décide d'informatiser une partie de ses opérations.

— **Traitement immédiat des résultats**: Aussitôt la dernière réponse fournie, on obtient un résultat qui peut par la suite être

manipulé comme toute autre donnée. On peut donc ajouter le résultat d'un étudiant à ceux d'un groupe, puis, avec les logiciels appropriés, réordonner, épurer, imprimer ces résultats pour produire des listes. On peut également les transformer par des formules mathématiques ou des regroupements, les analyser statistiquement ou les intégrer dans une banque de données.

— **Confidentialité des résultats**: Comme aucune feuille de réponses ne circule et que les résultats sont rangés dans un fichier, seul l'usager qui a accès à ce fichier peut connaître l'ensemble des résultats du moins tant qu'il ne décide pas de les imprimer ou de les copier ailleurs. Compain *et al.* (1989) ont montré que l'exigence de confidentialité pouvait parfois, à elle seule, justifier l'informatisation d'une épreuve, notamment quand l'évaluation se fait en milieu de travail.

— **Sécurité du test**: Non seulement il n'y a aucune feuille de réponse, mais il n'y aucun questionnaire. À moins qu'il ne s'agisse d'un test de certification où il faut observer le plus grand secret, on peut donc administrer le test sans risquer que les questions et les réponses ne soient divulgées. Compte tenu des coûts de production de versions équivalentes, il s'agit là d'un avantage appréciable. La sécurité du test est d'autant plus préservée qu'avec de grandes banques d'items, le contenu des tests peut varier considérablement d'un sujet à l'autre décourageant ainsi la production de copies illicites.

— **Administration individuelle**: Avec un test adaptatif, il n'est plus nécessaire de former des groupes afin de justifier la location d'une salle et l'emploi d'un surveillant. On peut procéder à des administrations ponctuelles: il suffit de placer l'étudiant devant sa machine. On peut même envisager une administration à distance en utilisant des lignes téléphoniques ou des réseaux d'ordinateurs.

— **Temps d'administration réduit**: Parce qu'il faut moins d'items pour atteindre des niveaux de précision comparables, l'administration se fait plus rapidement qu'avec une version conventionnelle. De plus, Greaud et Green (1986) font remarquer

que les sujets répondent plus rapidement à un test sur ordinateur du fait qu'au lieu d'inscrire la réponse sur une feuille, il n'ont qu'une touche à appuyer. Du point de vue de la gestion du temps dans le système d'éducation, il faut voir cette réduction du temps consacré à l'évaluation comme du temps supplémentaire pour des activités d'apprentissage encadrées.

– **Continuité avec l'enseignement programmé**: Dans la perspective d'un enseignement assisté par ordinateur, on peut imaginer le test de classement adaptatif comme la première étape d'un programme où les applications pédagogiques de l'ordinateur sont intégrées à un programme d'activités pédagogiques adaptées à l'étudiant.

5.2 Les limites du testing adaptatif

5.2.1 *Au plan psychométrique*

Les objections théoriques qu'on peut apporter au testing adaptatif sont de taille. Elles tiennent à la fois aux modèles psychométriques basés sur la théorie du trait latent et à la nature des tâches qu'on peut demander au sujet.

– **Unidimensionalité**: Dans notre discussion sur la dimensionalité des tests de langue, nous avons établi qu'il existait entre les composantes d'un test de langue une variance commune et qu'il était, de ce fait, possible de concevoir un test de maîtrise générale. Toutefois, l'élaboration d'un test adaptatif qui viserait à évaluer divers aspects de la performance linguistique, comme peut prétendre le faire un test diagnostique par exemple, pose des problèmes majeurs. À moins de postuler, comme les tenants des approches naturelles (Krashen 1978), une séquence naturelle d'acquisition des éléments linguistiques, on voit mal comment un instrument basé sur la théorie du trait latent peut servir à déceler des forces ou des faiblesses chez un étudiant. On peut donc douter de la valeur d'un test adaptatif dans le cadre d'une évaluation formative dont le but n'est pas tant de situer l'élève sur un continuum que de dégager des éléments permettant l'élaboration d'objectifs d'apprentissage

spécifiques. Par ailleurs, même en admettant comme Henning *et al.* (1985) que la procédure de calibration soit suffisamment robuste pour traiter un ensemble d'items d'un test de langue présentant une structure multidimensionnelle, on ne peut pas nécessairemnt conclure que l'application de la théorie soit légitime. En effet, toutes les applications faisant appel à des banques d'items s'appuient sur le principe d'invariance des items c'est-à-dire sur l'hypothèse que les items sont interchangeables. Tout item présentant un écart par rapport à l'axe commun devient donc une source d'erreur. Il s'agit là d'une restriction très sérieuse qui peut limiter singulièrement la comparabilité des résultats et conséquemment la valeur d'un test adaptatif qui ne respecterait pas l'exigence d'unidimensionalité.

— **Indépendance des items**: Aspect particulier de l'unidimensionalité, ce pré-requis à l'utilisation d'un modèle de la théorie du trait latent représente une condition que beaucoup de tests de langue ne satisfont pas. D'une part, en excluant l'emploi des réponses antérieures comme indices, le principe d'indépendance des items entrave le processus par lequel l'étudiant construit des hypothèses sur la signification d'un énoncé ou la formulation d'une réponse correcte et appropriée. D'autre part, elle interdit la réalisation de tâches purement intégratives, comme la production libre, en forçant une approche par item qui ne correspond pas toujours à ce qu'on doit mesurer.

— **Erreur de calibration**: Même avec des échantillons assez grands, Thissen et Wainer (1982) ont trouvé que la calibration avec les modèles à trait latent pouvait présenter des erreurs types assez considérables surtout quand la valeur des paramètres de difficulté et de discrimination (*a* et *b*) diminuait et que l'effet de hasard (*c*) augmentait. Comme la qualité de la mesure d'un test adaptatif repose sur la précision de la calibration, l'utilisation de paramètres inexacts peut invalider totalement la procédure.

— **Incompatibilité des modes de correction**: Si l'estimation du niveau d'habileté par maximum de vraisemblance s'avère la solution la plus juste du point de vue théorique, les problèmes inhérents à ce mode de correction forcent les praticiens à se tourner

vers d'autres méthodes. Il est douteux qu'on puisse comparer des résultats obtenus avec des modes de correction différents et la question de déterminer le mode le plus approprié reste ouverte. Comme le signalent Gialluca et Weiss (1979:26) en étudiant le déroulement d'un test adaptatif: *The issue of the appropriate choice of scoring method pervades implementations of ICC test theory and hence is not confined to this particular implementation of an adaptive testing strategy.* Le problème devient encore plus sérieux quand on veut comparer les résultats d'un test adaptatif avec ceux d'un test «papier-crayon» où on considère généralement le nombre de réponses correctes.

— **Comparabilité avec les versions conventionnelles**: La difficulté de comparer les résultats d'une version adaptatif avec ceux d'une version «papier-crayon» ne réside pas uniquement au plan du mode de correction. Comme le fait remarquer Green (1988), un item originalement conçu pour un test conventionnel prend, lorsque transposé sur un écran, un nouvel éclairage de sorte que la tâche que doit réaliser l'étudiant est différente. Quand la calibration des items se fait à partir des données d'une version «papier-crayon», il faut donc être prudent afin de limiter les interférences attribuables à la tranposition d'un mode de présentation à l'autre.

— **Interaction limitée**: On est encore loin du jour où l'étudiant pourra échanger avec la machine. Dans la plupart des cas, la réception des messages se fait au moyen d'un écran et la production au moyen d'un clavier. Afin de réduire les interférences du médium, il est habituellemnt préférable de limiter l'intervention de l'étudiant à l'utilisation de quelques touches. De plus, à un moment où on s'interroge encore sur la façon de donner une interprétation sémantique convenable à des échantillons de langue naturelle, on peut imaginer que l'analyse automatisée des productions libres à des fins évaluatives n'est pas pour demain. C'est pourquoi, pour l'instant, le testing adaptatif, convient davantage à l'évaluation des habiletés réceptives. De plus, dans l'optique cognitiviste (Anderson 1985), on peut affirmer que de par la nature des tâches qu'on peut faire réaliser, le testing adaptatif se prête mieux à l'évaluation des tâches déclaratives qu'à celle des tâches procédurales.

— **Mode écrit**: Des contraintes technologiques tendent à confiner l'application du testing adaptatif au mode écrit. C'est encore au prix d'un raffinement technique souvent prohibitif qu'on peut fournir à l'étudiant un signal audio ou vidéo. Comme pour la majorité des applications pédagogiques de l'ordinateur, le mode écrit demeure donc pour l'instant le type d'interaction privilégié dans le cadre d'un test informatisé.

— **Environnement artificiel**: La situation de test, à savoir réagir à des questions à choix multiple provenant d'un ordinateur, présente évidemment peu de similarités avec une situation réelle d'utilisation de la langue seconde. On peut de la sorte mesurer certains aspects composant la maîtrise générale et sous-jacents à l'utilisation effective de la langue seconde dans diverses situations. Il est cependant bien difficile de viser la mise au point d'un test direct ou même la réalisation de tâches pouvant mener à des généralisations vers des situations plus authentiques.

5.2.2 *Au plan psychologique*

Quels sont les effets de l'environnement d'un test adaptatif auprès des étudiants. De fait, on peut anticiper beaucoup d'inconvénients en réfléchissant sur les réactions habituelles que peut susciter l'emploi de l'ordinateur dans le milieu de l'éducation.

— **Manque de validité apparente**: Le cliché le plus commun a trait à l'aspect déhumanisant de la machine. Beaucoup se demanderont comment on peut imaginer communiquer avec un ordinateur. Ainsi, le meilleur test informatisé risque toujours d'être taxé de tels jugements pré-conçus. Ces stéréotypes font partie de la validité apparente du test. Il ne s'agit plus de savoir ce que le test mesure effectivement mais plutôt ce qu'il semble mesurer pour les usagers. Nevo (1985) considère que la validité apparente mérite une certaine attention et qu'elle est mesurable. Morrow (1979:155) accorde une place prépondérante à cette forme de validité allant même jusqu'à prétendre que la fiabilité est secondaire, *subordinate to face validity*". S'opposant vivement à cette position, Stevenson

(1985) rejette quant à lui toute forme de jugement naif sur la validité d'un test. Légitimes ou non, il reste que ces jugements s'observent et qu'on ne peut espérer voir un étudiant accepter un instrument qu'il juge négativement. Tous les sujets qui opposent ce type de résistance à l'emploi de l'ordinateur auront beaucoup de mal à se sentir à l'aise au cours d'une séance de testing adaptatif.

— **Environnement menaçant**: Pour beaucoup de sujets, l'environnement informatique peut constituer une source de crainte et d'anxiété. Pour ceux qui n'ont jamais touché à un clavier d'ordinateur ou même à un clavier de machine à écrire, la perspective de devoir utiliser un ordinateur peut être traumatisante quand elle s'ajoute à celle de devoir faire un test. La présence de la machine devient donc une source de «bruit» (Bowen 1978) qui peut empêcher l'étudiant de donner sa pleine mesure. On peut aussi penser que certains étudiants, les plus jeunes ou les fanatiques des machines, seront favorisés par la procédure au détriment des plus âgés ou des profanes des ordinateurs. Selon son style d'apprentissage, un étudiant peut être plus ou moins à l'aise face à un ordinateur: par exemple, Chapelle et Jamieson (1982) rapportent que l'enseignement assisté par ordinateur convient peu aux étudiants utilisant des stratégies d'apprentissage de type analytique.

— **Familiarisation avec l'environnement**: Alors que la situation classique du test objectif représente rarement une situation nouvelle pour l'étudiant, il faut, dans le cas d'une épreuve informatisée, s'assurer que chacun puisse communiquer avec la machine. Dans les cas les plus simples, comme celui du système *CAPT*, il peut s'agir essentiellement de repérer les touches sur le clavier et, dans les cas les plus compliqués, d'apprendre à appeler le programme ou de taper des commandes particulières. Dans tout test informatisé, il faut donc ajouter une composante didactique pour expliquer au sujet comment utiliser l'appareil.

— **Aucune révision**: Lorsqu'une réponse est entrée, elle l'est de façon irrémédiable. L'étudiant ne peut pas corriger sa réponse, pas plus qu'il ne peut revoir les questions antérieures. Cette particularité peut se justifier d'un point de vue psychométrique car,

dans un test de langue, la première réponse est probablement celle qui rend le mieux compte de la performance réelle de l'étudiant. Par contre, certains étudiants pourraient se sentir frustrés de ce que les stratégies de révision qu'ils mettent en oeuvre habituellement (relecture et retours) soient devenues inopérantes.

— **Durée variable**: Au cours d'un test conventionnel, l'étudiant peut généralement estimer où il se trouve dans le déroulement temporel du test en se guidant sur la feuille de réponses ou le temps écoulé. Avec un test informatisé adaptatif, où souvent le nombre d'items peut varier et où l'on n'établit pas de limite de temps, il est possible que le sujet ait l'impression d'être perdu dans la structure du test.

5.2.3 *Au plan administratif*

Aussi séduisant soit-il du point de vue administratif, un système de testing adaptatif pose des problèmes particuliers du point de vue de son implantation et de son maintien.

— **Temps et coûts de l'élaboration**: Il nous a fallu plusieurs années de travail pour mettre au point le système *CAPT* et nous ne pouvons toujours pas prédire s'il peut être utilisé avec une population différente de celle qui a servi à la calibration. La conception du test, la rédaction des items, la cueillette des données, la saisie des données, l'analyse des items, la calibration et la programmation sont des étapes nécessaires exigeant des ressources humaines et financières qui ne sont pas toujours disponibles. Comme il faut prévoir un échéancier assez long et un financement adéquat, il semble que le testing adaptatif se prête davantage à des utilisations à grande échelle.

— **Taille de l'échantillon**: Même dans les situations où on peut compter sur temps et argent, il faut encore s'assurer de disposer d'un nombre de répondants suffisant. Bien sûr, on peut réduire la taille de l'échantillon en utilisant des modèles à un seul paramètre (Lord 1983), mais comme le démontre Divgi (1986), ces

modèles risquent de ne pas être à la hauteur. Or, avec un modèle à trois paramètres, il est recommandé de soumettre les items à un millier de personne. De plus, pour mettre sur pied, une banque plus étendue, il faut prévoir un plan d'ancrage dont l'application peut s'avérer très lourde.

– **Coût de l'équipement**: Le système *CAPT* utilise des appareils de type IBM-PC à configuration minimale. Bien que le coût de ces appareils soit beaucoup moindre maintenant que lorsqu'ils sont apparus sur le marché, il faut non seulement prévoir le coût initial à l'achat ou les frais de location mais aussi l'allocation de l'espace, l'entretien et la surveillance. Dans le cas où l'on désire administrer plusieurs tests à la fois, il faudra prévoir plus d'un appareil. Notons également que dans certains établissements, ces appareils ne sont tout simplement pas disponibles et que la multiplicité des marques et des configurations ne fait qu'aggraver la situation.

– **Administration individuelle**: Si l'administration individuelle peut être avantageuse dans certains cas, elle devient problématique lorsqu'il faut évaluer beaucoup d'étudiants simultanément. L'utilisation d'un laboratoire de micro-ordinateurs peut résoudre ce problème mais on y trouve rarement plus d'une vingtaine de postes de travail et les coûts d'utilisation peuvent être très élevés.

– **Fiabilité de l'équipement**: Comme tout autre appareil, le micro-ordinateur n'est pas infaillible, surtout dans les conditions qui prévalent dans les milieux scolaires. Il faut donc prendre en considération la menace de panne, que ce soit à la suite d'un défaut ou d'un bris de l'appareil, d'une mauvaise installation ou d'une interruption de courant.

– **Fiabilité du logiciel**: L'élaboration de logiciels à des fins spécifiques implique souvent que les mises à l'essai ne peuvent pas toujours détecter certains problèmes de programmation. Il n'est donc pas impossible qu'avec tel équipement et tel utilisateur, le logiciel faille à la tâche.

⑥

DONNÉES COMPARATIVES EXPÉRIMENTALES

Contrairement au chapitre précédent, ce dernier chapitre a une orientation plus expérimentale. Abordant successivement les plans psychométrique (les effets sur la qualité de la mesure) et psychologique (les effets sur l'attitude et le comportement de l'étudiant), nous passerons en revue les études comparatives que nous avons recensées puis nous ferons le compte rendu de nos propres expérimentations. Au plan administratif, nous verrons comment, en prenant en considération les conditions dans lesquelles s'est déroulée notre expérimentation, on peut envisager l'implantation d'un test adaptatif dans le milieu de l'enseignement post-secondaire.

6.1 Le plan psychométrique

6.1.1 *Revue des études comparatives*

Il y une trentaine d'années, certains chercheurs s'intéressaient déjà aux utilisations de l'ordinateur en testing et tentaient de dégager les avantages psychométriques d'une administration informatisée. Linn *et al.* (1969) passent en revue les travaux effectués dans les années '60 sur les tests à branchement. En en comparant différents types, ils sont eux-mêmes amenés à conclure en la supériorité d'un test informatisé construit autour de la technique d'échantillonage séquentiel de Wald (1947). Sachar et Fletcher (1978) rappellent les expériences d'administration de tests informatisés du début des années '70 et se livrent eux-même à une comparaison au terme de laquelle ils concluent

que le fait d'administrer le même test de façon conventionnelle ou informatisée n'amène pas de changements notables du point de vue statistique.

Cherchant à vérifier le principe de l'invariance des items, Lord (1977b) simule les résultats qu'obtiendraient des étudiants avec un test «papier-crayon» fixe et unique pour les comparer avec ceux qu'ils obtiendraient avec un test composé d'items choisis au hasard par l'ordinateur. Bien que les différences soient assez ténues, il conclut néanmoins que la procédure conventionnelle permet de mieux rendre compte des différences individuelles. Par contre, dans une autre étude (Lord 1977c), ce même auteur observe qu'en simulant l'administration d'une épreuve d'aptitude verbale adaptative (c'est-à-dire tenant compte cette fois du niveau du sujet), on obtient un niveau de précision équivalent à celui d'une procédure conventionnelle et ce, avec deux fois moins d'items.

C'est au groupe de recherche du laboratoire de testing adaptatif de l'Université du Minnesota qu'on doit les études comparatives les plus poussées et les plus systématiques. Regroupés autour de David Weiss, ces chercheurs ont tenté, vers la fin des années '70, de préciser les avantages psychométriques du testing adaptatif. Ainsi Bejar *et al.* (1977) et Bejar (1978) ont essayé d'appuyer les positions théoriques du groupe sur le testing adaptatif avec des sujets réels. Il ont comparé les courbes d'information de la version adaptative d'un test de biologie comprenant cinq parties avec les courbes d'information de la version conventionnelle. La version adaptative utilisait une procédure *stradaptive* et une estimation de l'habileté par maximum de vraisemblance. Les chercheurs ont trouvé qu'avec la version adaptative, il suffisait de 27 items pour atteindre le degré de précision du test conventionnel original de 35 items. Ils ont aussi construit un test conventionnel qui regroupait les 25 items les plus discriminants de la banque pour trouver qu'avec un test adaptatif on atteignait une précision comparable avec seulement 17 items. Ces conclusions ont amené les chercheurs à s'interroger aussi sur la validité de ces deux versions du test de biologie (Bejar et Weiss 1978). En analysant la grille de corrélations obtenue entre deux formes (pré-test et post-test) de chaque version, ils ont observé

que les tests conventionnels et les tests adaptatifs mesuraient le même construit mais que ces derniers étaient plus valides puisque la composante attribuée l'erreur de mesure étaient moins importante. Toutefois, en ajoutant un test mesurant les aptitudes verbales, ils ont constaté que les corrélations étaient plus fortes avec le test adaptatif, ce qui laissait croire que la compréhension des explications sur la consigne du test adaptatif pouvait avoir un effet sur les résultats. Weiss et Brown (1978) ont aussi comparé les formes adaptatives et conventionelles d'un test de connaissances techniques comprenant 12 sous-tests. Cette fois-ci, ils utilisaient une correction bayésienne. Ils ont trouvé que les courbes d'information étaient à peu près identiques mais que les versions adaptatives employaient deux fois moins d'items. Par la suite, Gialluca et Weiss (1979) ont tenté de déterminer dans quelle proportion l'efficacité du test adaptatif dépendait de la procédure de sélection des items par rapport à la stratégie de passage d'un sous-test à l'autre. Il apparaissait clairement que même en utilisant le mieux possible les résultats des sous-tests précédents pour déterminer l'item de départ du sous-test suivant, la stratégie de passage d'un sous-test à l'autre jouait un rôle marginal. C'était donc la capacité du test adaptatif de rapidement sélectionner un item conforme au niveau réel du sujet qui expliquait le gain réalisé par rapport aux tests conventionnels.

Par la suite Kingsbury et Weiss ont fait porter les intérêts du groupe vers les tests de «maîtrise» c'est-à-dire les tests de certification où il s'agit de déterminer si le candidat a atteint un seuil de passage pré-établi. Abandonnant définitivement la procédure *stradaptive* et optant pour la correction bayésienne, les chercheurs (Kingsbury et Weiss 1979) ont construit des tests adaptatifs dans le domaine de la mécanique. L'expérience consistait à comparer ces tests avec des versions conventionnelles, en faisant varier le seuil de passage. Il s'est avéré que les décisions prises sur la base des résultats pour chaque version concordaient dans plus de 95% des cas, mais que les tests adaptatifs utilisaient de 30% à 60% moins d'items. Kingsbury et Weiss (1980a, 1983) se sont aussi demandé si le cadre de la théorie du trait latent était la seule avenue pour ces tests de certification. Il ont simulé trois types d'administration: le

test conventionnel (à correction bayésienne et selon le nombre de réponses exactes), le test adaptatif (à correction bayésienne) et un test appliquant la technique d'échantillonnage séquentiel de Wald (1947). Les deux derniers types se sont révélés plus efficaces. La technique d'échantillonage séquentiel était supérieure dans la situation, peu vraisemblable, où les items avaient tous des paramètres identiques. Autrement, la procédure adaptative permettait d'arriver à la décision la plus juste avec un minimum d'items. Ces résultats ont été confirmés par une autre étude de Kingsbury et Weiss (1981), cette fois avec des sujets réels faisant un test de certification en biologie: non seulement la version adaptative utilisait 80% moins d'items que la version conventionnelle, mais elle permettait aussi de réduire le taux d'erreurs de classification. Ces résultats sont d'autant plus probants qu'on avait choisi les items de la version conventionnelle de façon à cibler l'épreuve autour du seuil de passage.

Le groupe de recherche de l'Université du Minnesota s'est aussi intéressé à vérifier la fiabilité et la validité concurrente des tests adaptatifs. Kingsbury et Weiss (1980b) ont administré à 472 étudiants universitaires, deux tests de vocabulaire conventionnels de 30 questions à choix multiple de même que deux tests de vocabulaire adaptatifs utilisant des banques différentes mais équivalentes. Il ont constaté que pour atteindre le degré de fiabilité inter-formes des tests conventionnels, il suffisait de 10 items avec la procédure adaptative. Enfin Martin *et al.* (1983) ont voulu établir de façon claire et définitive la supériorité des tests adaptatifs au plan psychométrique. Ils ont administré à plus de 250 recrues de la marine américaine, deux tests conventionnels d'aptitude verbale comprenant chacun 30 items. Les sujets ont aussi fait un test témoin conventionnel de 50 items puis deux tests adaptatifs (à correction bayésienne), construits à partir de deux banques regroupant des items différents mais de même nature. Trois conclusions ressortaient de leur étude:

- pour atteindre une fiabilité inter-formes de .8 il fallait administrer 17 items avec la procédure conventionnelle mais seulement 9 avec la procédure adaptative;

- il suffisait de seulement 4 items pour que la variance entre les résultats des formes équivalentes adaptatives cesse d'être significative alors qu'il en fallait 14 pour les formes conventionnelles;
- les tests adaptatifs montraient un indice de validité (la corrélation inter-formes après correction) supérieur, surtout avec peu d'items, le gain en validité devenant négligeable après l'administration du quinzième item.

Les recherches du groupe de l'Université du Minnesota relativement à la validité ne sont néanmoins pas tout à fait satisfaisantes. Elles révèlent surtout que la procédure adaptative peut éliminer plus rapidement une partie de l'erreur de mesure mais n'écartent pas entièrement la possibilité d'une variance spécifique à chacun des types de tests. Green *et al.* (1984) soulignent que le mode de présentation peut avoir un effet sur le construit et même le contenu d'un test. Par exemple, Biskin et Kolatchin (1977) constatent, dans la mise au point d'un test de personnalité, que le nombre d'omissions peut varier d'une version à l'autre et qu'il faut donc s'assurer que le fait qu'il suffise d'appuyer sur une touche n'incite pas les sujets à annuler plus facilement une réponse. Par contre, Green (1988) examine les corrélations entre les diverses parties d'une batterie de tests d'aptitude administrés de façon conventionnelle et selon une procédure adaptative; il trouve que la structure factorielle ne varie pas et conclut que le construit est identique. Qu'en est-il des tests de langue? Canale (1981b) affirme qu'un effet de méthode est toujours susceptible de se manifester dans un test de langue et reconnaît cinq variables: le mode (écrit ou oral), le type de réponse, la procédure d'administration, l'environnement (physique et affectif) et le mode de correction. Shohamy (1984) par exemple, rapporte que le type de réponse (à choix multiple plutôt qu'ouverte ou en L1 plutot qu'en L2) affecte les résultats d'un test de lecture. Il est certain qu'on ne peut pas toujours isoler la variable responsable de l'effet de méthode mais il est possible de vérifier l'importance de cet effet.

6.1.2 *Comparaison entre les administrations*

Au plan psychométrique, notre expérimentation visait à établir des comparaisons entre les différentes versions de notre propre test afin, d'une part, de vérifier la comparabilité des résultats par la validité inter-formes et, d'autre part, de comparer la fiabilité c'est-à-dire la marge d'erreur des versions l'une par rapport à l'autre. Contrairement à plusieurs recherches que nous avons rapportées, nous nous en sommes tenu non seulement à des réponses réelles mais aussi à des instruments dont la construction réflète les contraintes pratiques qui jouent dans l'acceptabilité d'un test de classement en langue seconde. Ainsi, en ce qui concerne les versions «papier-crayon», leur longueur (60 items) correspond à un maximum acceptable pour ce type de test. La correction de ces versions se fait en comptant le nombre de réponses correctes puisque ni une correction bayésienne ni une correction par maximum de vraisemblance ne sont envisageables dans la pratique. De plus, nous ne nous intéressions pas tant au score qu'au niveau auquel ce score correspondait dans l'échelle que nous avons préalablement établie. Pour ce qui est des versions informatisées, nous ne considérions que le niveau final et ce bien que nous soyons conscient que la procédure de correction puisse varier selon que l'estimation par maximum de vraisemblance a réussi ou non. En d'autre termes, nous avons cherché à reproduire la situation qui se présente effectivement quand il s'agit de choisir quelle version est la plus appropriée. Nous n'avons donc pas essayé de minimiser la composante «méthode»; au contraire, nous nous préoccupions d'en déterminer l'importance.

6.1.2.1 *Administrations simulées*

L'unité de développement du système *CAPT* offre la possibilité de simuler des séances de testing adaptatif soit par stratification (STRAT), soit par correspondance (MATCH). Pour les sujets qui ont répondu à tous les items de la banque (l'échantillon d'analyse de la version 2), il est donc possible de prédire le résultat à une version informatisée. Par ailleurs, les résultats des sous-ensembles

que représentent les versions 3.1 et 3.2 étaient toujours disponibles. Nous avons donc retiré 50 feuilles de réponses de l'échantillon d'analyse de façon à obtenir une distribution à peu près équivalente des quatorze niveaux de maîtrise. Nous n'avons pas retenu les sujets qui n'avaient pas complété les épreuves ou ceux dont la configuration de réponses avait été jugée inadéquate par rapport au modèle.

TABLEAU 6.1
Répartition de niveaux selon l'habileté

Valeur	Niveau	Habileté	Score
01	Vrai débutant	-4.0 ~ -1.2	de 0 à 21
02	Vrai débutant +	-1.2 ~ -0.88	de 22 à 25
03	Faux débutant	-0.88 ~ -0.6	de 26 à 28
04	Faux débutant +	-0.6 ~ -0.4	de 29 à 31
05	Intermédiaire I	-0.4 ~ -0.2	32 et 33
06	Intermédiaire I+	-0.2 ~ -0.05	34 et 35
07	Intermédiaire II	-0.05 ~ 0.05	36 et 37
08	Intermédiaire II+	0.05 ~ 0.2	38 et 39
09	Intermédiaire III	0.2 ~ 0.4	40 et 41
10	Intermédiaire III+	0.4 ~ 0.6	de 42 à 44
11	Avancé	0.6 ~ 0.88	de 45 à 47
12	Avancé +	0.88 ~ 1.2	de 48 à 51
13	Très avancé	1.2 ~ 2.0	de 52 à 59
14	Très avancé +	2.0 ~ 4.0	60

Le test informatisé distingue quatorze niveaux soit les sept niveaux que nous avons déjà définis, plus six niveaux mitoyens (les cas frontières) et un niveau supérieur (les scores parfaits). La division doit théoriquement partager la population en un nombre égal de sujets. Nous avons remanié la table de conversion que nous avions établie pour les versions «papier-crayon» (cf tableau 3.14). L'équivalence avec les scores bruts s'est effectuée selon la même technique que celle qui avait servi à la première division c'est-à-dire de façon à minimiser les écarts entre les versions 3.1 et 3.2. Nous avons alors établi le niveau auquel ces sujets auraient été classés en ne considérant que les items retenus pour les versions 3.1 et 3.2. Afin de pouvoir calculer les corrélations entre le classement auquel conduit

chaque type de tests, on a assigné ensuite une valeur numérique à chaque niveau. La répartition des quatorze niveaux s'établit selon le tableau 6.1.

Nous avons ensuite procédé à 50 simulations de séances de testing adaptatif par stratification en transcrivant ce qui apparaissait sur la feuille de réponse au numéro correspondant à l'item présenté. Nous avons répété l'opération avec la procédure par correspondance.

6.1.2.1.1 Les corrélations entre les formes

Le tableau 6.2 montre les moyennes et les écarts types obtenus avec chaque test. Il apparaît que les deux versions informatisées sous-estiment l'habilité des étudiants. Certains étudiants pourraient être classés à un niveau plus bas par rapport au classement des tests conventionnels (versions 3.1 et 3.2). Le test *t* pairé (avec *SPSS-PC*) confirme par ailleurs que les écarts entre les tests utilisant le même mode de présentation ne sont pas significatifs ($p > .1$). Par contre, les écarts entre la version par stratification et l'une ou l'autre des versions «papier-crayon» le sont ($p < .005$).

TABLEAU 6.2
Moyennes et écarts types des simulations

Version	Moyenne	Ecart-type
3.1	7.08	4.194
3.2	7.34	4.429
STRAT	6.50	4.032
MATCH	6.68	4.058

n = 50

Notons qu'étant donné que les versions 3.1 et 3.2 sont parallèles, il n'est pas étonnant que l'écart entre leur moyenne

et leur variance ne soit pas significatif. À cet égard, la corrélation entre les deux versions conventionnelles peut être considérée comme un indice de la fiabilité entre les deux formes, une fois réalisée la conversion des scores bruts en niveaux de maîtrise.

TABLEAU 6.3
Corrélations entre les versions

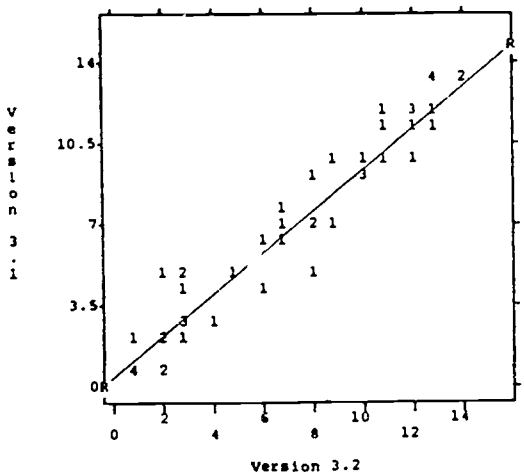
	3.1	3.2	STRAT	MATCH
3.1	-	.969	.947	.907
3.2	.969	-	.945	.915
STRAT	.948	.945	-	.932
MATCH	.907	.915	.932	-

n = 50, $p < .001$

L'examen de la grille des coefficients de corrélation du tableau 6.3 révèle que malgré la différence de moyenne, le test adaptatif par stratification montre une corrélation relativement élevée avec les deux versions conventionnelles (r >.94). Par contre, les corrélations avec la version par correspondance sont plus faibles. Il faut rappeler que l'erreur acceptable qui servait de critère d'arrêt, était plus grande, dans la version MATCH. Il se peut donc que la différence de fiabilité explique le fait que les coefficients impliquant la version par correspondance soient moins élevés.

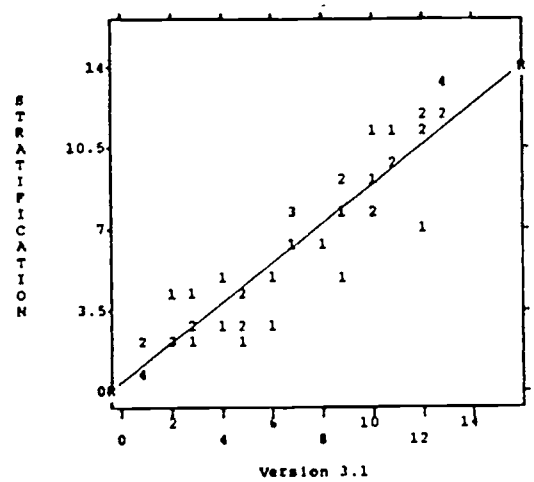
Le diagramme de dispersion de la figure 6.1 montre que, comme prévu, les résultats des versions 3.1 et 3.2 pour les 50 sujets retenus, se concentrent autour de la ligne de régression. Les diagrammes des figures 6.2 et 6.3 montrent comment se comparent le classement obtenu avec la version par stratification et celui obtenu avec les versions conventionnelles. Les points suivent de près la ligne de régression mais s'en écartent entre les niveaux 8 à 12 (intermédiaires forts). Dans cette région, le classement du test par stratification serait moins sûr que celui des versions conventionnelles.

FIGURE 6.1
Diagramme de dispersion des versions 3



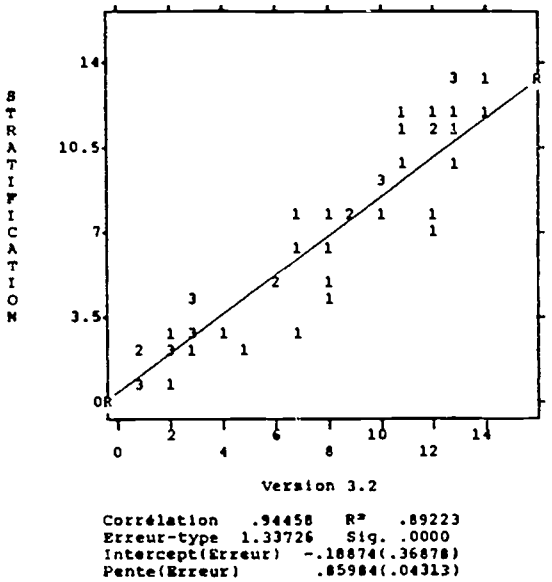
Corrélation .96874 R² .93846
Erreur-type 1.05111 Sig. .0000
Intercept(Erreur) .34768(.28986)
Pente(Erreur) .91721(.03390)

FIGURE 6.2
Versions 3.1 vs Test par stratification



Corrélation .94755 R² .89786
Erreur-type 1.30189 Sig. .0000
Intercept(Erreur) .05004(.36400)
Pente(Erreur) .91101(.04435)

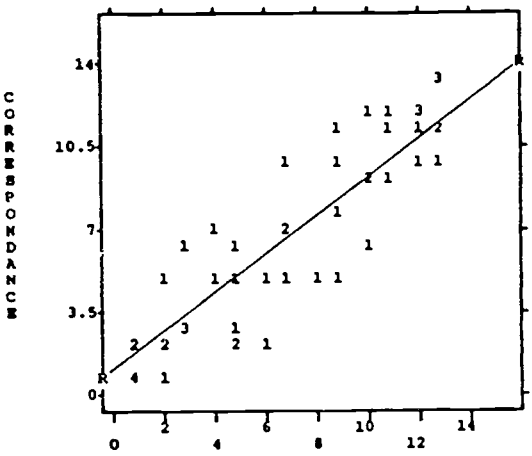
FIGURE 6.3
Versions 3.2 vs Test par stratification



Il est beaucoup plus difficile de localiser les différences entre le test adaptatif par correspondance et les versions conventionnelles. En examinant les diagrammes des figures 6.4 et 6.5, il est clair que les points s'écartent considérablement de la ligne de régression. Cela signifie qu'il y a des écarts importants entre les deux évaluations de certains sujets: jusqu'à quatre niveaux dans certains cas! Il semble bien que l'erreur de mesure dans le cas de cette procédure soit trop grande pour conduire à des décisions quant au niveau d'un étudiant. Il n'est donc pas étonnant que le regroupement des points soit assez diffus dans le diagramme de la figure 6.5 où l'on compare les deux tests informatisés. Il convient toutefois de noter que comme les moyennes des résultats fournis par ces deux types de tests sont similaires, la pente de la ligne de régression se rapproche de 1 (soit un angle de 45°) et l'intercept se rapproche de 0. Il est donc permis de croire que n'eût été de l'erreur de mesure, la corrélation entre les deux tests informatisés aurait été plus forte. Enfin, il faut noter qu'il semble y avoir une meilleure correspondance entre les deux procédures aux niveaux extrêmes de l'échelle.

FIGURE 6.4

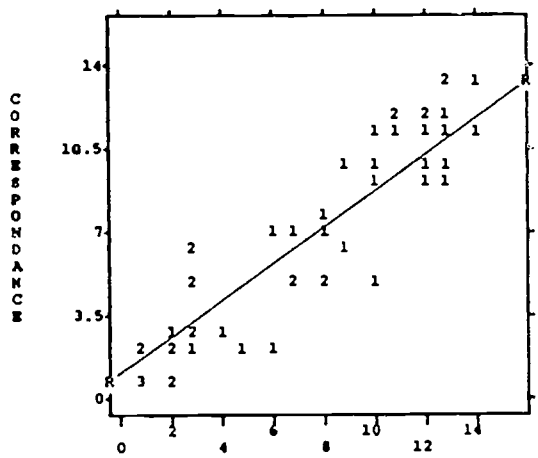
Versions 3.1 vs Test par correspondance



Corrélation .90700 R² .82264
Erreur-type 1.72668 Sig. .0000
Intercept(Erreur) .46602(.48277)
Pente(Erreur) .87768(.05882)

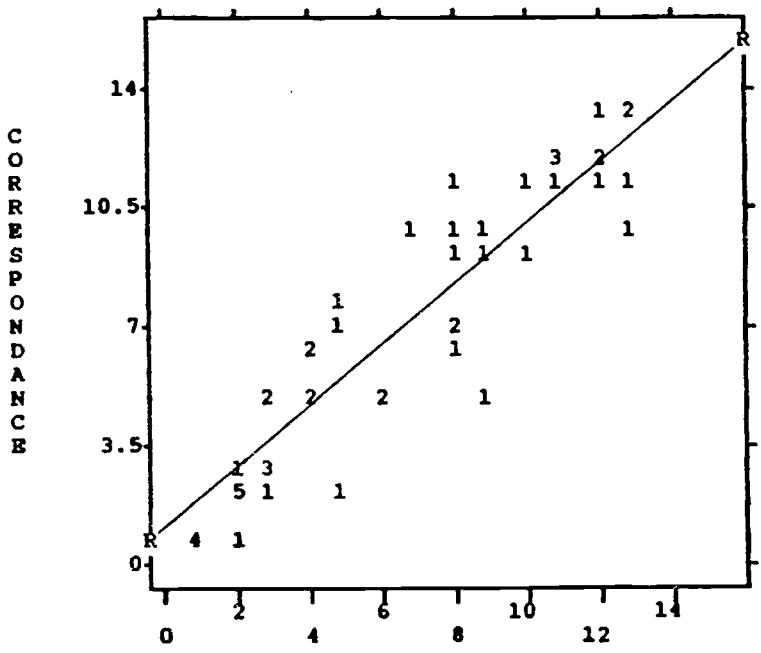
FIGURE 6.5

Versions 3.2 vs Test par correspondance



Corrélation .91457 R² .83664
Erreur-type 1.65815 Sig. .0000
Intercept(Erreur) .52956(.45727)
Pente(Erreur) .83794(.05348)

FIGURE 6.6
Versions informatisées (STRAT vs MATCH)



Corrélation	.93180	R^2	.86825
Erreur-type	1.48817	Sig.	.0000
Intercept(Erreur)		.58395(.40221)	
Pente(Erreur)		.93785(.05273)	

Le fait que les corrélations soient toutes supérieures à .9 pourrait laisser croire qu'on peut interchanger les résultats de ces quatre types de tests. Toutefois, il faut garder à l'esprit que la transformation des scores bruts en un nombre limité de catégories et le fait que les tests mesurent une large gamme d'habileté sont deux facteurs qui tendent à gonfler les coefficients de corrélation. Il faut donc aussi prendre en considération le nombre d'accords et de désaccords par rapport au nombre de décisions à prendre pour se rendre compte de l'effet des écarts entre les résultats des quatre types de tests. Le tableau 6.4 indique le nombre de décisions où les niveaux de classement ne concordent pas.

TABLEAU 6.4
Nombre de désaccords sur le niveau

	Version 3.2	STRAT	MATCH
Version 3.1			
1 niveau ou +	29 (58%)	32 (64%)	31 (62%)
2 niveaux ou +	8 (16%)	10 (20%)	19 (38%)
3 niveaux ou +	2 (4%)	4 (8%)	10 (20%)
MATCH			
1 niveau ou +	35 (70%)	32 (64%)	
2 niveaux ou +	18 (36%)	13 (26%)	
3 niveaux ou +	12 (24%)	6 (12%)	
STRAT			
1 niveau ou +	35 (70%)		
2 niveaux ou +	11 (22%)		
3 niveaux ou +	6 (12%)		

On voit que dans la majorité des cas (entre 58% et 70%), le niveau auquel un étudiant serait classé avec un test donné ne correspond pas au niveau auquel il serait classé avec un autre test. Il est évident qu'aucun de ces instruments n'est assez précis pour permettre un classement satisfaisant avec 14 niveaux. Par contre, comme dans la plupart des programmes de langue, on ne distingue pas plus de sept niveaux, ce sont plutôt les écarts de deux niveaux ou plus qui peuvent causer des problèmes. Quant aux écarts de plus de trois niveaux ou plus, ils devraient être exceptionnels. Dans cette perspective, seules les versions 3.1 et 3.2 seraient interchangeables et pourraient, par exemple, servir à mesurer le progrès réalisé sur une année. On note de nombreuses divergences entre les résultats obtenus avec la procédure MATCH et les autres résultats. Pourtant, les deux tests adaptatifs donnent aussi des résultats assez proches bien qu'ils pourraient difficilement satisfaire les exigences de tests parallèles.

6.1.2.1.2 Les courbes d'information

Green (1984) signale que les indices de la théorie classique s'avèrent peu utiles pour déterminer la fiabilité des tests adaptatifs car ces indices portent sur l'ensemble des sujets. Il propose un

indice de fiabilité marginale et un indice de fiabilité conditionnelle en précisant qu'il s'agit d'approximations plus ou moins justes. De fait, dans le cadre de notre étude comparative, il semble beaucoup plus pertinent d'examiner les courbes d'information. En effet, nous connaissons déjà la fiabilité relative des deux versions «papier-crayon» puisque les courbes d'information ont servi à construire ces deux versions parallèles.

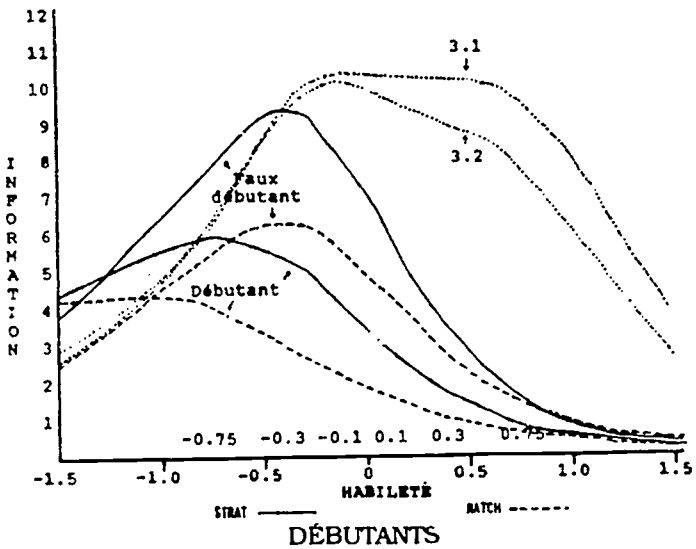
Nous voulions connaître l'information que pouvaient fournir les procédures adaptatives pour chacun des sept niveaux principaux. Pour chacun des modes d'administration, nous avons d'abord éliminé, parmi les 50 simulations, celles qui avaient mené à un classement dans une catégorie mitoyenne (Vrai débutant + ... Très avancé +). En évitant d'utiliser une bande d'habilité trop large, nous estimions donner une idée plus juste de la fiabilité à un niveau particulier. Exploitant la propriété d'additivité de la fonction d'information, à l'aide d'une variante du programme TICC, nous avons fait la somme de l'information obtenue à chacun des sept niveaux, pour chaque sous-test dont le résultat final correspondait au niveau considéré. Nous avons ensuite fait les moyennes de façon à pouvoir tracer la courbe d'information pour sept sujets typiques, représentant chacun un niveau, de «Débutant» jusqu'à «Très avancé». On peut ainsi visualiser la fiabilité des tests adaptatifs selon les niveaux, par rapport aux versions 3.1 et 3.2.

Les figures 6.7a, 6.7b et 6.7c montrent les courbes d'information pour le sous-test #1 (compréhension). Chez les vrais et les faux débutants (figure 6.7a), on voit que dans les deux versions du test adaptatif, la courbe d'information culmine non pas au niveau auquel le programme a évalué le sujet, mais au niveau supérieur. On peut expliquer ce phénomène de deux façons. D'une part, quand on s'approche des niveaux extrêmes de l'échelle, l'estimation de l'habileté du sujet ne se situe plus au centre de l'intervalle de confiance. En d'autres termes, un débutant n'aurait pas pu être classé plus bas que le niveau débutant mais aurait pu être classé à un niveau supérieur de sorte qu'on doit utiliser un certain nombre d'items pour écarter cette dernière hypothèse. D'autre part, on trouve peu d'items très faciles dans la banque; le programme doit donc recourir à des items plus difficiles pour arriver à accumuler

l'information nécessaire. On peut donc penser que l'addition d'items ayant un paramètre de discrimination élevé et un paramètre de difficulté très bas permettrait de mieux cibler le test adaptatif au niveau des débutants. Le nombre relativement grand d'items présentés réflète cette lacune. Chez les débutants et les faux débutants, il a fallu en effet une moyenne de 11.36 items (n =11) avant l'arrêt de la procédure avec le programme STRAT et 10.7 items (n = 11) avec le programme MATCH. Notons par ailleurs que la procédure par stratification permet d'obtenir, au niveau «Faux débutant», beaucoup plus d'information que la version par correspondance ou que les deux versions conventionnelles. Au niveau «Débutant», les deux tests adaptatifs sont plus précis que les versions conventionnelles.

FIGURE 6.7a

Courbes d'information du sous-test #1



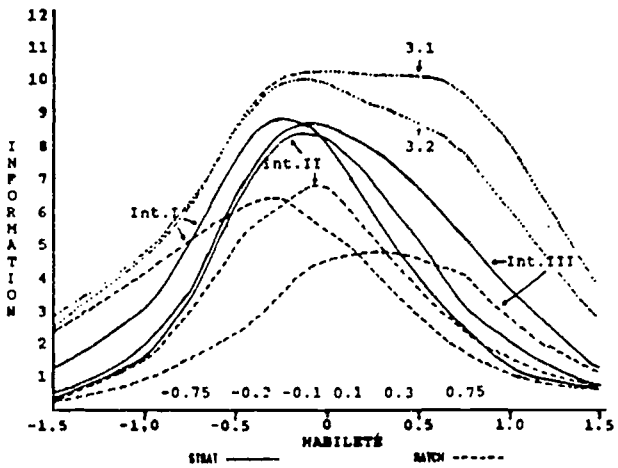
DÉBUTANTS

Avec les niveaux intermédiaires, (figure 6.7b), on voit clairement que les versions conventionnelles sont beaucoup plus précises. Cependant, avec une moyenne de seulement 7.33 items (n =9), la procédure STRAT permet d'obtenir un résultat d'une précision étonnante. Par contre, la procédure MATCH donne

beaucoup moins d'information même si elle utilise plus d'items (moyenne de 8.27, n =11). Cela tient au fait que la procédure par stratification tend à ne présenter que les items les plus discriminants. C'est pour cette raison que nous avons dû établir un seuil d'erreur acceptable moins élevé pour la procédure par correspondance qui, quant à elle, utilise généralement une plus grande variété d'items. Par ailleurs, il ne faut pas donner trop d'importance au fait qu'avec la procédure de stratification, le test de niveau «Intermédiaire III» semble plus facile que celui du niveau «Intermédiaire II»; cela s'explique du fait qu'on ne trouvait qu'un seul sujet classé à ce niveau. De même, la forme plus aplatie du test pour «Intermédiaire III» avec la procédure par correspondance peut être attribuée au nombre limité de sujets à ce niveau. D'ailleurs, il s'agit là d'un problème sérieux qui s'est manifesté au cours des expérimentations: bien que la théorie prédisait une distribution à peu près égale du nombre de sujets entre les différents niveaux, il y a beaucoup moins de représentants aux niveaux intermédiaires (de -0.5 à 0.5). C'est donc chez les intermédiaires qu'on risque de trouver le plus grand nombre de désaccords entre les classements obtenus à partir des différents tests.

FIGURE 6.7b

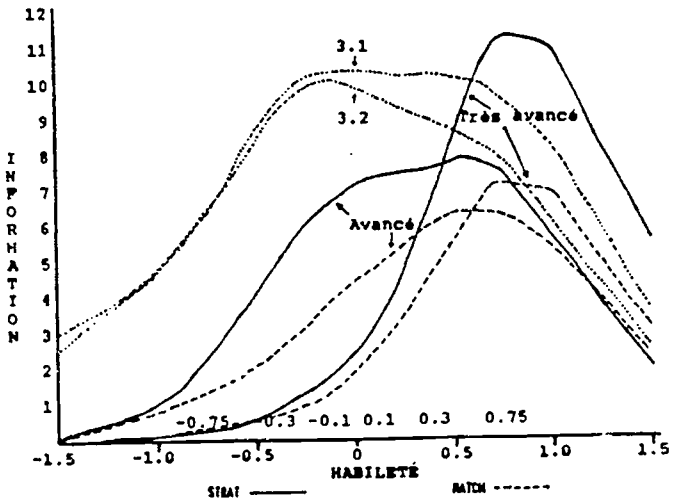
Courbes d'information du sous-test #1



INTERMÉDIAIRES

Dans la figure 6.7c, on observe, particulièrement chez les sujets de niveau «Très avancé» avec STRAT, un phénomène inverse à celui que nous avons noté chez les débutants. En effet, la courbe d'information culmine au niveau précédant celui auquel est estimée l'habileté des sujets. Les mêmes raisons peuvent être invoquées: déplacement par rapport à l'intervalle de confiance et manque d'items aux niveaux extrêmes. D autre part, ce qui frappe davantage chez les avancés (niveaux «Avancé» et Très avancé»), c'est que l'information obtenue avec les quatre types de tests varie peu. On peut donc dire qu'à ces niveaux, la fiabilité est comparable. Cependant, alors que les versions conventionnelles utilisent 20 items, le programme STRAT n'en utilise que 9 en moyenne (n =8) et le programme MATCH que 8 (n =8).

FIGURE 6.7c
Courbes d'information du sous-test #1



AVANCÉS

Le tracé des courbes pour le deuxième sous-test (énoncé approprié) qu'on trouve dans les figures 6.8a, 6.8b et 6.8c, illustre un des avantages du testing adaptatif. Lors de la création de la version «papier-crayon», nous avions constaté certains problèmes reliés à la faible discrimination de l'ensemble des items de cette

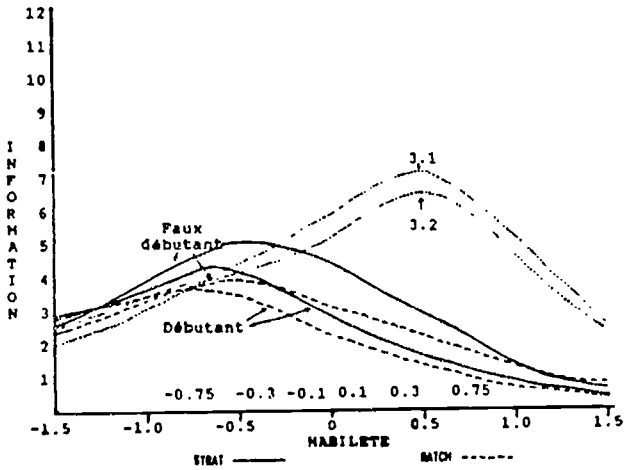
partie par rapport à ceux des autres sous-tests. Plusieurs items des versions 3.1 et 3.2, n'apportent qu'une contribution marginale à la mesure; quelques-uns peuvent même, pour une habileté donnée, diminuer la fonction d'information. Dans une telle situation, la procédure adaptative permet de ne présenter que les items susceptibles de réduire l'erreur de mesure pour un niveau donné.

On constate ainsi que pour le sous-test #2, les deux procédures adaptatives permettent d'obtenir sensiblement la même précision en utilisant beaucoup moins d'items.

Le nombre d'items utilisés par les tests adaptatifs varie peu d'un niveau à l'autre ou d'une procédure à l'autre: Avec STRAT, on utilise en moyenne 11.24 items (n =21) et avec MATCH, 11.32 (n =25). Ces nombres dépassent toutefois ce qu'on trouve pour les procédures adaptatives des autres sous-tests. Étant donné, la faible discrimination de certains items, on atteint souvent le maximum possible de 12 items. La procédure est alors interrompue avant que l'on ait atteint le niveau d'information visé. Enfin, comme pour le premier sous-test, on peut observer que le point maximum des courbes d'information pour les niveaux extrêmes (débutants et avancés) se déplace vers le centre de l'échelle.

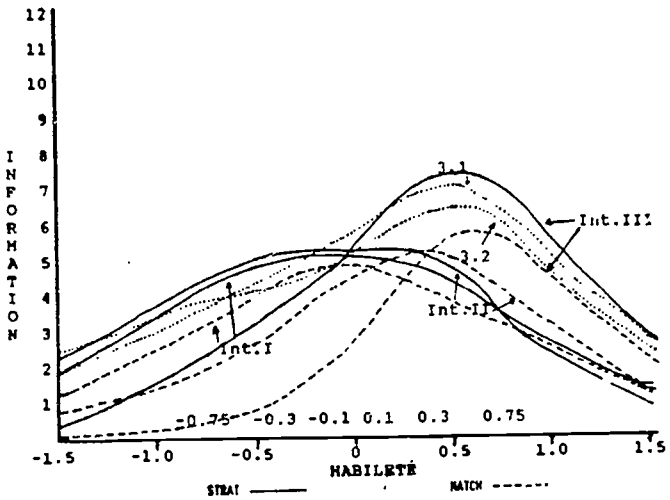
FIGURE 6.8a

Courbes d'information du sous-test #2



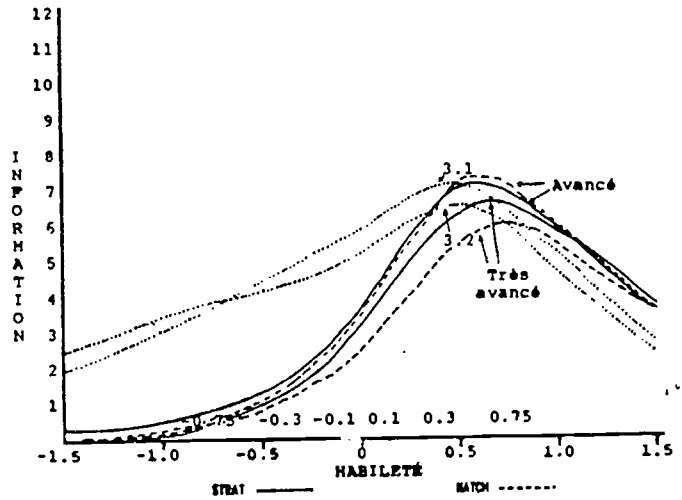
DÉBUTANTS

FIGURE 6.8b
Courbes d'information du sous-test #2



INTERMÉDIAIRES

FIGURE 6.8c
Courbes d'information du sous-test #2



AVANCÉS

Les courbes d'information du sous-test #3 (phrases à trou) montrent une situation où le test adaptatif permet, avec moins d'items, d'obtenir une fiabilité non seulement égale mais même supérieure à celle d'une version «papier-crayon». Le nombre d'items requis par la procédure par stratification varie peu d'un niveau à l'autre: la moyenne est de 10 items (n =28). Pourtant avec deux fois moins d'items, le test STRAT permet de recueillir plus d'information que les versions conventionnelles que ce soit chez les débutants (figure 6.9a), chez les intermédiaires (figure 6.9b) ou chez les avancés (figure 6.9c). On est donc loin de la distribution rectangulaire de la fiabilité que prédit la théorie. Il faut cependant noter que la fiabilité étonnante qu'on observe au niveau «Intermédiaire II» pourrait être le produit du hasard puisqu'avec STRAT, ce niveau n'était représenté que par un seul sujet.

La procédure par correspondance, quant à elle, utilise un peu moins d'items, soit une moyenne de 9.28 (n =29) pour l'ensemble des niveaux. Néanmoins, on recueille beaucoup moins d'information qu'avec la procédure par stratification. Par rapport à cette dernière et par rapport aux tests conventionnels, qui apportent nettement plus d'information au niveau «Intermédiaire III», on trouve avec MATCH une distribution plus rectangulaire de la fiabilité. Ainsi, la fiabilité est plus élevée chez les débutants et les avancés. Par contre, elle tend à être moins grande chez les sujets intermédiaires. Enfin, il convient de rappeler qu'ici encore, on peut observer la centralisation des courbes des niveaux extrêmes vers le centre de l'échelle.

En conclusion, l'examen des trois courbes d'information nous montre donc que malgré la centralisation des courbes d'information aux niveaux débutants et avancés, les tests adaptatifs qui s'adressent aux sujets appartenant à ces niveaux sont habituellement deux fois plus courts que les test conventionnels et leur précision égale (avec la procédure par correspondance) ou supérieure (avec la procédure par stratification). Par contre, aux niveaux intermédiaires, les items supplémentaires que comptent les versions conventionnelles contribuent à rendre le test plus précis sauf pour les sous-tests #2 et #3 administrés par stratification. Le gain en fiabilité avec les versions 3.1 et 3.2 au niveau intermédiaire peut être important (par exemple, par rapport au sous-test #1 par correspondance) ou négligeable (par exemple, par rapport au sous-test #2 par correspondance).

FIGURE 6.9a
Courbes d'information du sous-test #3

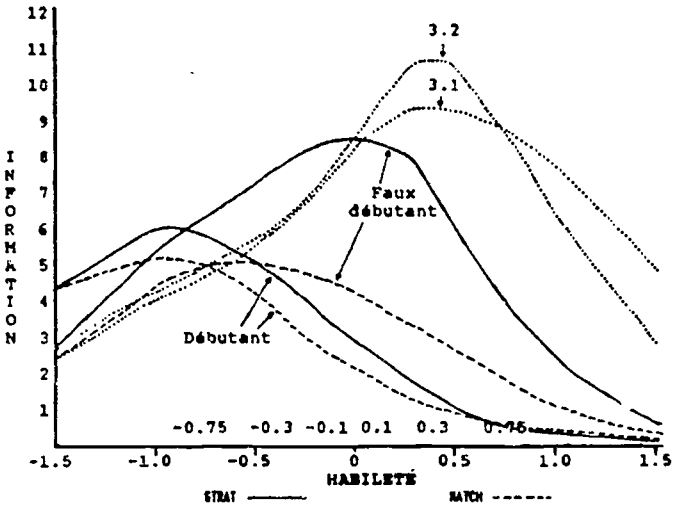


FIGURE 6.9b
Courbes d'information du sous-test #3

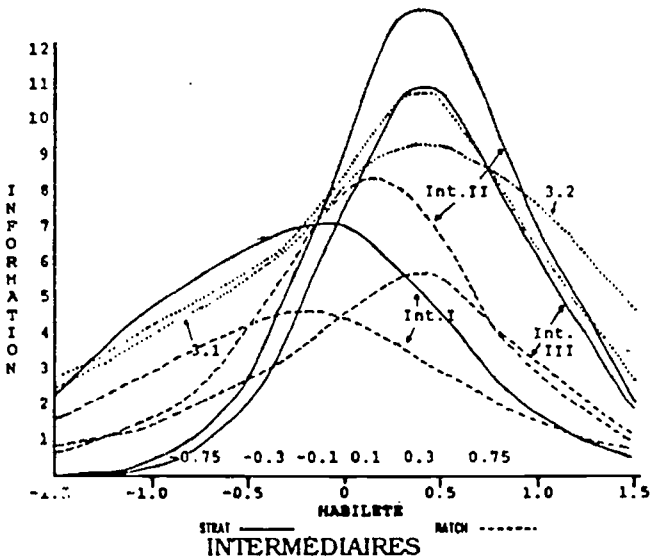
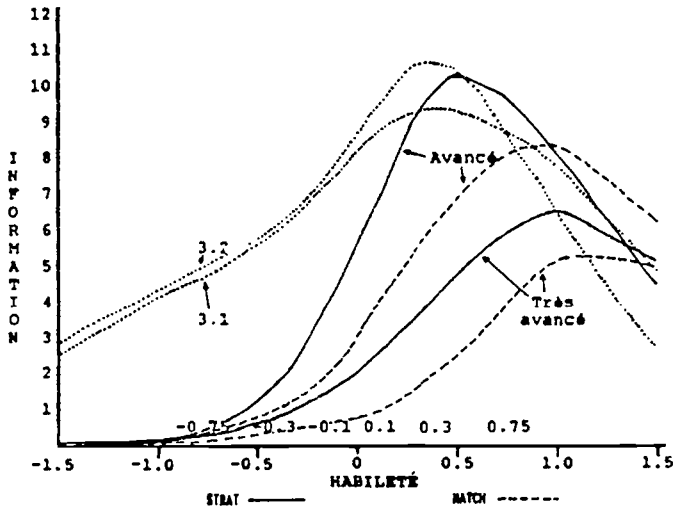


FIGURE 6.9c
Courbes d'information du sous-test #3



AVANCÉS

6.1.2.2 *Administrations expérimentales*

La supériorité de la procédure par stratification par rapport à la procédure par correspondance tient principalement au fait que le programme STRAT construit une grille où sont ordonnés pour chaque niveau, les dix items les plus susceptibles d'apporter de l'information. Ce ne sont donc que les items les plus discriminants qui seront présentés. La procédure MATCH donne plus d'importance au paramètre de difficulté de sorte que les items varient davantage d'un sujet à l'autre. Toutefois, avec une banque comportant peu d'items, dont certains par surcroît montrent un paramètre de discrimination tout juste acceptable, la procédure MATCH risque de demander plus d'items pour réduire l'erreur à la même marge.

Étant donné la composition de la banque, nous avons choisi le programme STRAT pour l'expérimentation du test informatisé dans une situation authentique. Nous cherchions à confirmer

l'effet de méthode que l'analyse précédente avait décelé. Cet essai a eu lieu au printemps 1988. Une centaine d'étudiants boursiers inscrits à l'Université York à Saint-Georges devaient faire un test au début et à la fin d'un programme intensif de six semaines. Une partie d'entre eux ont fait le test informatisé comme pré-test (groupe 1) tandis que l'autre partie a fait la version 3.1 (groupe 2). Les sujets du groupe 1 étaient en grande partie des étudiants déjà inscrits à l'Université York, pour qui le pré-test ne risquait pas de changer l'attribution de cours. Il n'y a aucune raison de croire que cette division, fort pratique, ait pu être moins adéquate qu'une division purement aléatoire. À la fin du programme, les étudiants ont reçu, en guise de post-test, la version qu'il n'avaient pas reçue au début: STRAT pour ceux qui avait fait la version 3.1 et vice-versa. À cause de retards et de départs prématurés du programme, certains étudiants n'ont fait qu'un type de test; nous n'avons retenu que les 83 étudiants qui avaient fait les deux types tests. La version conventionnelle se donnait dans deux salles de classe et on avait prévu environ une heure et demie; la version informatisé se donnait dans un laboratoire de micro-ordinateurs par sous-groupes qui changeaient aux demi-heures.

Le tableau 6.5 montre les moyennes obtenues par chaque groupe à la suite des deux administrations. Ainsi que nous l'avions relevé lors des simulations, les étudiants qui avaient reçu la version conventionnelle au pré-test (groupe 2) semblent avoir été favorisés.

TABLEAU 6.5
Pré-test et post-test (St-Georges)

		n	Moyenne	Écart-type
PRÉ-TEST	Ensemble	83	7.325	3.351
	Groupe 1	47	6.936	3.674
	Groupe 2	36	7.833	4.286
POST-TEST	Ensemble	83	9.157	3.225
	Groupe 1	47	9.447	3.133
	Groupe 2	36	8.778	3.348

Les résultats au post-test ne sont pas aussi clairs car ils réflètent à la fois les progrès réalisés par les apprenants et l'effet du type de test administré.

Soit y_{ij1} : Résultats du pré-test du sujet j du groupe i
y_{ij2} : Résultats du post-test du sujet j du groupe i
d_{ij1} : Différence entre le pré-test et le post-test

$$d_{ij} = y_{ij2} - y_{ij1}$$

$$j = 1, \ldots n_i \; ; \quad i = 1, 2 \; .$$

Afin de neutraliser la variation entre les sujets nous utilisons une analyse de variance à mesures répétées en utilisant le modèle suivant:

$$d_{ij} = \mu + \alpha_i + e_{ij}$$

où $\sum_{i=1}^{2} \alpha_i = \alpha_1 + \alpha_2 = 0$ et $e_{ij} \sim N(0, \sigma^2)$

On vérifie deux hypothèses: 1) $H_0 : \mu = 0$
2) $H_0 : \alpha_1 = \alpha_2 = 0$

TABLEAU 6.6
Analyse de variance: pré-test et post-test

Source de variation	Somme des carrés	Degrés de liberté	Moyenne des carrés	F	Signification de F
A l'intérieur des cellules	337.63	81	4.17		
CONSTANTE	243.35	1	243.35	58.38	.000
GROUPE	50.00	1	50.00	12.00	.001

La procédure *MANOVA* de *SPSS-PC* nous fournit les résultats du tableau 6.6. La vérification de la constante (μ) sert à examiner la validité de la première hypothèse nulle qui est clairement rejetée du fait que F = 58.38, ce qui est très élevé (p = .000). Cette première observation est encourageante pour les organisateurs puisqu'elle a trait à l'efficacité du programme. Toutefois, elle nous intéresse peu sinon pour confirmer la validité du test et la pertinence de la

division initiale à sept niveaux. Il est en effet raisonnable qu'en un programme intensif de six semaines (ou un programme régulier d'une année scolaire) les meilleurs étudiants aient progressé d'un niveau.

La vérification de la variable «GROUPE» revêt un plus grand intérêt car elle permet de vérifier la deuxième hypothèse, celle qui concerne l'ordre des tests c'est-à-dire le mode d'administration. Comme $F = 12$, il s'agit d'une source importante de variation. Avec un résultat aussi significatif ($p < .001$), on confirme, lors d'une administration réelle, la présence dans le test conventionnel et le test informatisé (STRAT), d'un effet de méthode que nous avions observé avec les simulations. Cela implique qu'à moins de pouvoir isoler la variation attribuable à l'instrument, on ne peut pas utiliser indifféremment l'une ou l'autre des versions pour mesurer l'efficacité d'un programme. Cela implique aussi que dans le contexte de l'expérimentation, il est très difficile de mesurer les progrès individuels ou de comparer les sujets entre eux.

On peut s'interroger sur les causes de cet effet de méthode. On peut penser que le mode de présentation (écran vs questionnaire) serait responsable de la variation. On peut aussi penser que le mode de réponse (clavier vs feuille de réponse) contribuerait à ce facteur de méthode. Toutefois, il nous semble que c'est plutôt du côté du mode de correction qu'il faut chercher la réponse. Les versions «papier-crayon» utilisent un nombre fixe d'items, toujours les mêmes d'un sujet à l'autre, et sont habituellement corrigées en comptant le nombre de réponses exactes. Bien sûr, ce score peut éventuellement être transformé, notamment en le normalisant (score z). Les versions adaptatives utilisent un nombre variable d'items, par définition différents d'un sujet à l'autre, et sont corrigées électroniquement par des techniques qui tiennent compte de la configuration des réponses à partir des paramètres des items. Le fait d'exprimer les résultats des versions adaptatives en termes d'écarts par rapport à la courbe normale ne garantit pas nécessairement une équivalence stricte avec les scores norma-

lisés des versions conventionnelles. La recherche d'une échelle commune pour les tests conventionnels et adaptatifs est un problème complexe auquel nous n'avons pas trouvé de solution tout à fait satisfaisante et qui rend difficile toute comparaison entre les deux types de tests. Il est possible que des études plus approfondies sur ces aspects statistiques puissent mener à une solution à l'incompatibilité des modes de sélection. Il faut néanmoins reconnaître que ces études pourraient aussi mettre en cause la pertinence de la théorie du trait latent dans ce genre d'applications.

6.2 Le plan psychologique

6.2.1 *Les études comparatives*

Outre les caractéristiques psychométriques, il faut considérer l'aspect psychologique. À quoi bon mettre au point un instrument de mesure extrêmement précis si ses effets sur l'apprenant sont désastreux? Il faut donc se demander quel impact le mode d'administration peut avoir sur les attitudes et les comportements pendant le test. Par ailleurs, il est clair que les plans psychométrique et psychologique sont interreliés et que les effets psychoaffectifs peuvent éventuellement avoir des incidences sur la qualité de la mesure. Pourtant, ce qui nous intéresse ici ce ne sont pas tant les résultats en tant que tels mais plutôt les réactions des sujets c'est-à-dire les sentiments qu'éprouvent ces sujets face à l'instrument, les perceptions qu'ils en ont et les comportements qu'ils adoptent au cours du test. On sait que dans le cas d'un test de langue ces considérations sont primordiales, que les critères d'ordre purement docimologique ne suffisent pas (Shohamy 1982).

Il ne faut pas s'étonner des conclusions de Rushinek *et al.* (1985) selon lesquels les étudiants qui réussissent moins bien à un cours informatisé ont une perception plus négative de l'enseignement assisté par ordinateur. Par contre, dans le cadre du testing

adaptatif où la sensation d'échec devrait être moins fréquente qu'avec un test conventionnel, on peut se demander si cette observation est généralisable.

En faisant la revue des rares études sur les aspects psychologiques des tests informatisés, Koch et Patience (1978) font ressortir l'aspect anxiogène de ces tests. Les réponses d'étudiants de collèges américains à leur questionnaire d'attitude après l'administration d'un test adaptatif et d'un test conventionnel montrent que la préférence va vers le mode d'administration qui génère le moins d'anxiété: les étudiants préfèrent donc les tests conventionnels. Les étudiants motivés, c'est-à-dire ceux pour qui le résultat revêt une certaine importance, se disent plus anxieux devant un test informatisé et, par conséquent, l'apprécient moins. Par ailleurs, ceux qui sont moins motivés le trouvent plus difficile et de ce fait seraient également plus critiques face au testing adaptatif.

Ces observations sur l'anxiété et la motivation contrastent avec celles formulées par l'équipe de recherche de l'Université du Minnesota qui s'est aussi intéressée aux aspects psychologiques du testing adaptatif. Weiss et Betz (1976a, 1976b) ont en effet trouvé que si l'anxiété était plus grande lors d'un test adaptatif, la motivation l'était également. Les auteurs ont aussi constaté que le fait de connaître la bonne réponse sur le champ au cours d'une séance de test adaptatif, avait des effets positifs. Prestwood (1978) a nuancé ce jugement en affirmant que la rétroaction instantanée n'influait par sur le résultat comme tel mais qu'elle était néanmoins fort appréciée des étudiants. Pine (1978) et Pine *et al.* (1979) ont quant à eux corroboré non seulement l'augmentation du degré d'anxiété et de motivation avec le testing adaptatif mais aussi l'effet généralement positif de la rétroaction instantanée. Ils insistent toutefois sur le fait que la rétroaction instantanée a des effets négatifs chez les étudiants de race noire. En effet, ces derniers semblent mieux réussir aux tests adaptatifs qu'aux tests conventionnels sauf quand on leur présente la bonne réponse immédiatement. Par ailleurs, le fait que le testing adaptatif favorise les Noirs va dans le sens des

observations de Johnson et Mihal (1973) pour qui l'informatisation d'un test se traduisait par une amélioration des résultats chez les minorités noires.

6.2.2 *Analyse quantitative*

6.2.2.1 Le questionnaire

Il est assez courant d'utiliser un questionnaire afin de vérifier l'effet psychologique d'un type d'enseignement ou d'un test. Rushinek *et al.* (1985) ont utilisé un questionnaire pour connaître la réaction des apprenants à la suite d'un cours de formation informatisé. Watts (1989) a eu recours au même type de questionnaire pour évaluer l'impact d'un programme avec vidéodisque. Madsen (1982) mesure l'anxiété générée par divers types de tests de langue et souligne que ces données sont indispensables car cette variable n'affecte pas tous les apprenants de la même façon. Plusieurs questionnaires visant à mesurer l'attitude des étudiants face à l'entrevue d'évaluation de l'oral ont été mis au point et administrés (Shohamy 1982, Scott 1986, Zeidner et Bensoussan 1988). Notons que ces derniers questionnaires sont construits à partir d'échelles Likert; ils sont conçus avec l'intention de soumettre les réponses à une analyse factorielle en vue de réduire les données à un nombre minimal de facteurs. Ils visent donc à porter un jugement global sur l'attitude des apprenants relativement à un mode d'évaluation particulier.

Notre approche est quelque peu différente car nous avons plutôt cherché à confirmer un certain nombre d'hypothèses quant à la réaction des étudiants vis-à-vis les versions «papier-crayon» et informatisées de notre test. Ainsi comme chaque question est destinée à être analysée individuellement, le questionnaire se présente plutôt comme un sondage. Chaque question portant sur un aspect particulier, le questionnaire ne présente pas de redondances. On trouve le plus souvent trois cases (c'est-à-dire trois possibilités de réponse) mais certaines questions en comportent quatre ou cinq car il nous semblait plus important de fournir au répondant une gamme d'options signifi-

catives que de conserver la même échelle tout au long du questionnaire. Bien que ne se prêtant pas à une analyse statistique très fine, ce questionnaire nous semblait approprié pour mettre le doigt sur des contrastes dominants entre les deux types de tests.

Nous avons administré le questionnaire aux groupes avec lesquels nous avons procédé à la mise à l'essai du logiciel, au printemps 1988, à Saint-Georges (Université York). Rappelons qu'un groupe avait fait le test conventionnel comme pré-test (version 3.1) et que l'autre avait fait le test informatisé. Au post-test, six semaines plus tard on a inversé les tests. On demandait aux étudiants de remplir le questionnaire dès qu'ils avaient complété le test. Le tableau 6.7 résume la distribution des questionnaires recueillis.

TABLEAU 6.7
Répartition des questionnaires

Version	Pré-test	Post-test	Total
Conventionnelle	40	50	90
Informatisée	48	38	86
Total	88	88	176

6.2.2.2 *Les résultats*

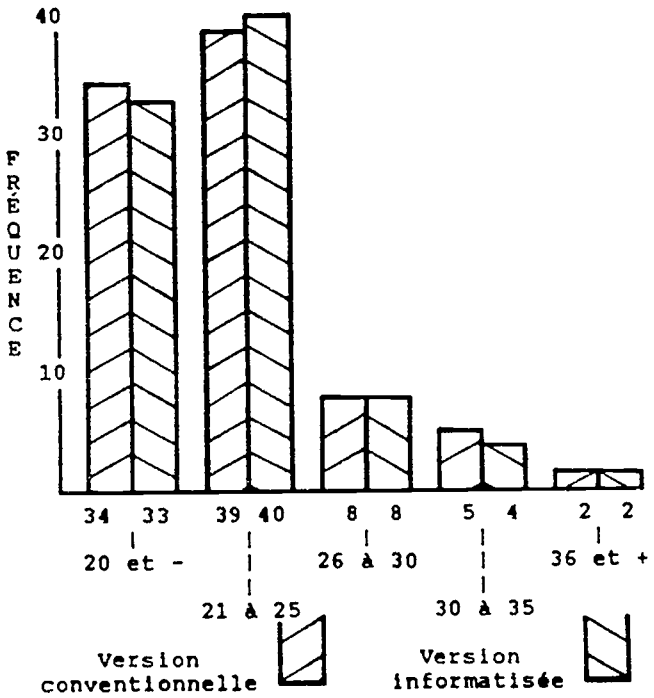
6.2.2.2.1 Les variables démographiques

Nous espérions faire des recoupements avec les caractéristiques que nous pensions pouvoir expliquer certaines différences dans les perceptions, les attitudes ou les comportements face au test: l'âge, le domaine d'études, la familiarité avec les ordinateurs et la langue maternelle.

Après s'être identifié, l'étudiant devait indiquer sa date de naissance. Il apparait que les deux groupes étaient sensiblement du

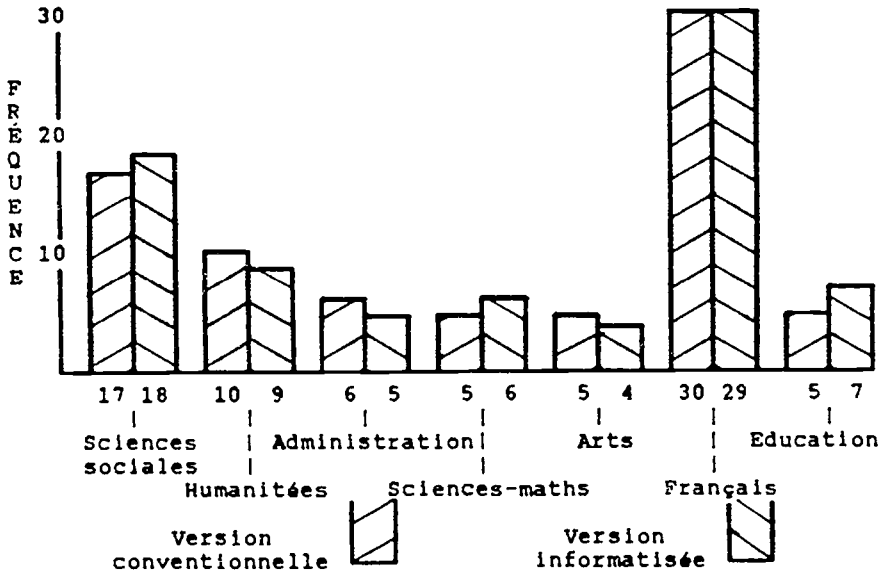
même âge, la moyenne s'établisssant à 23.14 ans pour les 88 sujets qui avaient fait la version 3.1 et à 23.16 ans pour ceux qui avait fait le test STRAT. La figure 6.10 montre la répartition selon les tranches d'âge que nous avons par la suite établies.

FIGURE 6.10
Répartition selon l'âge



La figure 6.11 montre la répartition de l'échantillon en fonction des domaines d'étude. Nous avons regroupé les programmes et spécialités en sept domaines distincts. L'étude des langues fait normalement partie des «Humanités» mais nous avons distingué les étudiants qui ont déclaré se spécialiser en français puisque que de par son grand nombre et ses intérêts particuliers, ce groupe constitue une catégorie en soi. Les deux échantillons restent tout à fait comparables.

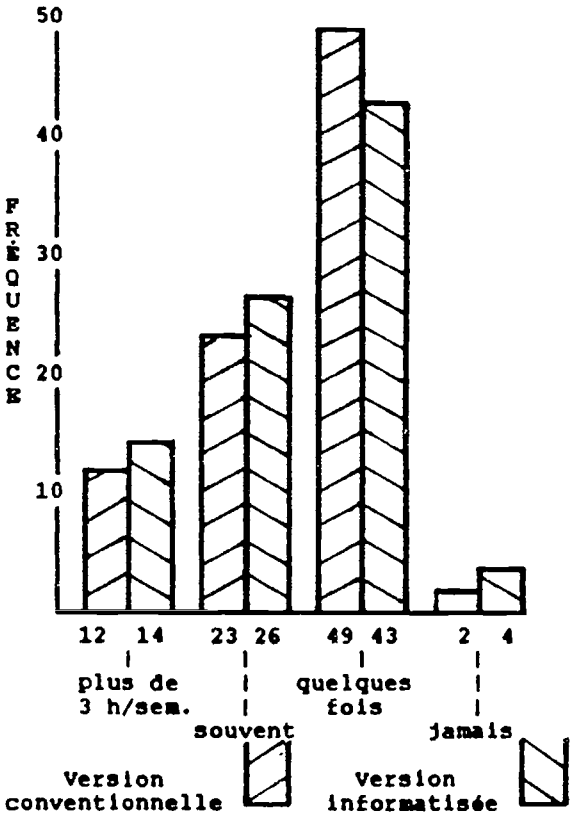
FIGURE 6.11
Répartition selon le domaine d'étude



Pour vérifier le degré de familiarité avec l'ordinateur, on demandait aux étudiants à quelle fréquence, ils utilisaient un ordinateur. En examinant l'histogramme de la figure 6.12, on s'étonne de constater que la majorité d'entre eux ont déclaré n'utiliser un ordinateur que de temps à autre. Quelques étudiants (2 avec la version conventionnelle et 4 avec la version informatisée) ont avoué toucher à un ordinateur pour la première fois.

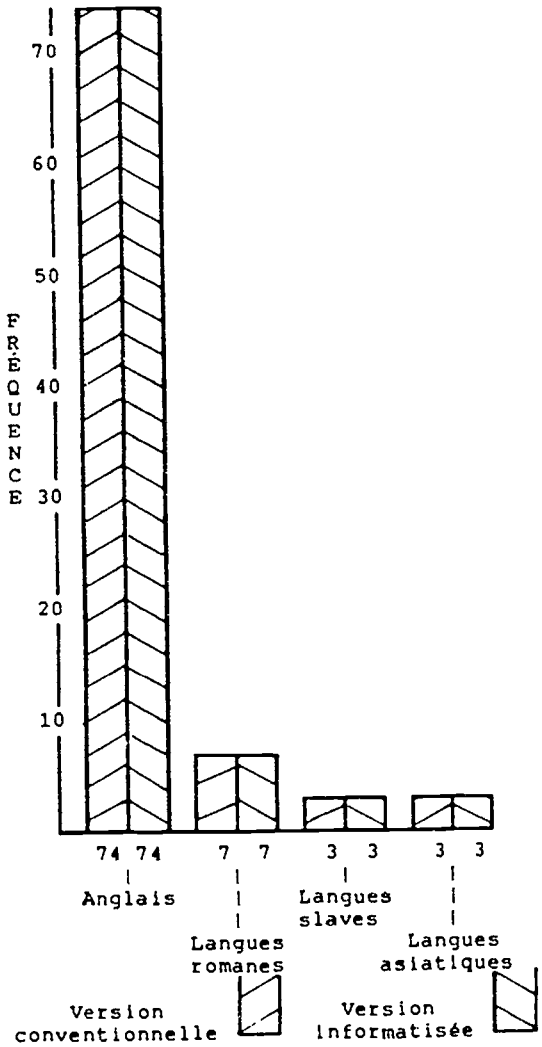
Par ailleurs, contrairement à ce qu'on pourrait penser en comparant ces réponses avec la variable «âge», on voit que les plus âgés ne sont pas nécessairement ceux qui ont moins d'expérience avec les ordinateurs. Par contre, en ce qui concerne le domaine d'études, les quelques étudiants inscrits dans des programmes de sciences-maths se démarquent clairement des autres car plus de 90% d'entre eux disent utiliser un ordinateur «souvent» ou «plus de trois heures par semaine». Par ailleurs, on ne note pas de différence appréciable entre les deux groupes de l'expérience.

FIGURE 6.12
Répartition selon la familiarité avec les ordinateurs



Enfin du point de vue de la langue maternelle, la dominance de l'anglais ressort nettement (figure 6.13). De ce point de vue l'échantillon est très homogène d'autant plus que les sujets qui déclarent une autre langue que l'anglais comme langue maternelle ont tous une bonne connaissance de l'anglais. Le programme de bourses s'adresse en effet à des étudiants non francophones, ayant le statut de citoyen canadien ou d'immigrant reçu et inscrits dans des établissements dispensant des cours en anglais. Cette variable n'est donc pas susceptible d'intervenir dans la compréhension de la consigne.

FIGURE 6.13
Répartition selon la langue maternelle



6.2.2.2.2 Les réactions au test

Les huit autres questions visaient à déterminer les réactions des étudiants face au test qu'ils venaient de faire.

— **Question #6:** *Comment évaluez-vous le degré de difficulté du test? Très facile, facile, juste, difficile ou très difficile?*

Comme le test adaptatif tient compte du niveau d'habileté de l'étudiant, on doit s'attendre à ce que les étudiants ne le trouvent ni difficile, ni facile. Par contre, les étudiants avancés devraient trouver le test conventionnel trop facile alors que les débutants devraient le trouver trop difficile.

Si on assigne une valeur numérique aux cinq choix de réponse proposés (de 1, pour «Très facile» à 5, pour «Très difficile»), les moyennes ne devraient pas être sensiblement différentes mais la variance devrait changer d'une version à l'autre. De fait, la moyenne des 88 sujets qui ont répondu au questionnaire après le test informatisé est de 2.89 et celle des 88 sujets qui ont répondu après le test conventionnel de 3.01. En appliquant le test *t* de *SPSS-PC*, on voit que l'écart n'est pas significatif ($p < .294$). En examinant la distribution des réponses du tableau 6.8, on voit que l'écart tient au fait que moins de sujets ont trouvé la version informatisée difficile ou très difficile; toutefois, comme le montre le test du chi carré ces différences sont peu importantes.

TABLEAU 6.8
Perception de la difficulté (#6)

	Très facile	Facile	Juste	Difficile	Très difficile	Total
3.1	3	17	47	18	3	88
STRAT	3	20	50	14	1	88
Total	6	37	97	32	4	176

$X^2 = 1.836$ $dl = 4$ $p = .766$

Aucun des recoupements que nous avons pu faire avec les variables socio-démographiques ne permet de dire qu'un groupe particulier aurait pu trouver une version plus ou moins difficile. Tout au plus peut-on dire que la trentaine de sujets qui se spécialisent en français ont trouvé les tests plutôt faciles (moyenne pour les deux

tests de 2.58); cette observation n'a rien de surprenant quand on sait que les spécialistes se retrouvent généralement parmi les groupes les plus avancés. Ce qui est surprenant c'est que cette tendance est plus marquée pour le test adaptatif (moyenne de 1.22) alors qu'on aurait pu s'attendre au contraire! Par ailleurs, comme le pré-test précède l'apprentissage, il est normal que les étudiants jugent la version conventionnelle de ce pré-test plus difficile: moyenne de 3.21 au pré-test et de 2.83 au post-test. Cependant, on s'étonne d'observer un phénomène semblable avec les tests adaptatifs dont la moyenne passe de 2.94 à 2.83. Il faut souligner néanmoins que l'écart est beaucoup moins marqué de sorte qu'on peut l'attribuer au hasard. Il reste qu'une conclusion s'impose: bien que la procédure du test adaptatif consiste à présenter des items qui sont d'un niveau correspondant à celui du sujet, celui-ci n'a pas nécessairement l'impression de faire un test sur mesure.

- **Question #7:** *Si ce test devait être utilisé pour vous classer dans un groupe correspondant à votre niveau de français, selon vous, où auriez-vous été classé? Au-dessus de votre niveau, au-dessous de votre niveau ou au niveau approprié?*

Le fait que les sujets soient peu conscients de ce que certains items trop difficiles ou trop faciles de la version «papier-crayon» mesurent mal leur niveau se réflète dans les réponses de la question #7 (tableau 6.9). En effet, tant pour la version conventionnelle que pour la version informatisée, 67 sujets, soit la grande majorité, considèrent que le test devrait les classer dans le groupe qui leur convient. Parmi les quelques-uns qui pensent le contraire, un plus grand nombre estiment être sur-évalués par le test «papier-crayon» alors qu'un plus grand nombre estiment être sous-évalués par le test adaptatif. Bien que peu significative, cette tendance peut s'expliquer par le fait que, comme nous l'avons observé lors des simulations, les versions conventionnelles tendent effectivement à sur-évaluer les étudiants.

TABLEAU 6.9
Niveau de classement (#7)

	Au dessus	En dessous	Au bon niveau	Total
3.1	13	8	67	88
STRAT	9	11	67	87
Total	22	19	134	175

$X^2 = 1.195$ $dl = 2$ $p = .55$

En faisant des recoupements avec les variables socio-démographiques, on note que parmi les 14 sujets dont l'anglais n'est pas la langue maternelle, aucun n'estime avoir été sous-évalué par le test et ce, quelle que soit la version administrée. On s'aperçoit aussi que ceux qui ne se spécialisent pas en français ont plus souvent l'impression que la version conventionnelle les a classés au-dessus de leur niveau. En effet, des 58 non-spécialistes, 16 se considéreraient mal classés; de ces 16, 11 se considéreraient classés à un niveau trop avancé. Cependant, il semble bien que, dans l'ensemble, les écarts sont minimes. Ainsi, le fait que les items soient choisis en fonction du niveau du sujet ne contribue pas à modifier la perception de ces sujets quant à l'exactitude de la décision relative à leur classement.

— **Question #8:** *Comment ce type de test mesure-t-il votre niveau général en français? Avec précision, assez bien ou mal?*

Compte tenu des résultats de la question précédente, on devrait s'attendre à ce que bon nombre de sujets accordent une cote élevée au test du point de vue de sa précision, que ce soit avec la version conventionnelle ou avec la version informatisée. Toutefois, en regardant la distribution des réponses du tableau 6.10, on peut s'étonner de ce que peu de sujets aient coché la case «Avec précision». En assignant une valeur numérique à chaque choix, on trouve que l'ensemble des étudiants juge le test «Assez bien» sans distinguer de quelle version il s'agit: 2.08 pour la version conventionnelle et

1.99 pour la version informatisée. Par contre, on voit que la version 3.1 est considérée comme mesurant mal deux fois plus souvent que la version par stratification. Il nous semble que cet écart s'explique non pas par la différence de fiabilité entre les versions mais plutôt par le fait qu'avec la version informatisée, les sujets jugent de la précision du test par le résultat que le programme leur a communiqué tandis qu'ils jugent de la précision de la version conventionnelle à partir du nombre de questions qu'ils croient avoir réussies.

TABLEAU 6.10
Précision du test (#8)

	Avec précision	Assez bien	Mal	Total
3.1	5	70	12	87
STRAT	7	73	6	86
Total	12	143	18	173

$X^2 = 2.391$ $dl = 2$ $p = .303$

Il semble que la différence se situe chez les plus jeunes (les 34 sujets âgés de moins de 21 ans). En effet, ceux-ci sont beaucoup plus critiques à l'égard de la version «papier-crayon» puisque que un sur cinq (7/34) juge qu'elle mesure mal.

— **Question #9:** *Que pensez-vous de la durée du test? Trop long, trop court ou bien?*

On sait que la procédure adaptative utilise deux fois moins d'items. Il est cependant intéressant de savoir si le test paraît trop long ou paraît trop court pour les étudiants. En observant les résultats du tableau 6.11, on est frappé par le fait que, malgré une différence appréciable entre la durée objective des deux types de tests, la grande majorité des étudiants ne les jugent ni trop longs ni trop courts. On peut penser que bien que peu sensibles au fait que la procédure adaptative choisisse les items en fonction de leur niveau, les étudiants reconnaissent qu'on ne peut évaluer la durée

du deux types de tests avec les mêmes critères. On observe néanmoins qu'une dizaine de sujets jugent le test informatisé trop court et qu'un nombre égal juge le test conventionnel trop long. Nous avons aussi remarqué que les étudiants semblaient plus indifférents à la durée lors du pré-test car ces divergences apparaissent surtout au post-test.

TABLEAU 6.11
Longueur du test (#9)

	Trop long	Trop court	Bien	Total
3.1	10	5	72	87
STRAT	2	10	75	87
Total	12	15	147	174

$X^2 = 7.061$ $dl = 2$ $p = .029$

La seule information qu'on puisse retirer des recoupements avec les variables socio-démographiques est reliée à la familiarité avec la machine. Bien que le nombre de sujets ne permette pas de conclusion définitive, il semblerait que plus l'étudiant utilise souvent un ordinateur, plus il estime que le test «papier-crayon» est trop long: 0 sur 2 (0%) chez ceux qui n'avaient jamais touché à un ordinateur, 5 sur 49 (10.2%) chez ceux qui en avaient utilisé un quelques fois, 3 sur 22 (13.6%) chez ceux qui en utilisent un souvent, 2 sur 12 (16.7%) chez ceux qui en utilisent un plus de trois heures par semaine.

- **Question #10:** *Dans un test, l'étudiant doit savoir ce qu'il doit faire. Les directives étaient-elles claires? Très claires, généralement claires ou ambigües?*

Dans les versions conventionnelles, chaque partie du test est précédée d'un exemple et on peut compter sur le fait que le test écrit à choix multiple est un type de test avec lequel les étudiants sont

familiers. De fait, lors de l'expérimentation nous n'avons jamais eu à répondre à des questions ou à fournir des explications supplémentaires concernant la consigne du test «papier-crayon». Avec la version informatisée, il nous a fallu prévoir dans la programmation, des explications sur la façon de répondre et sur le fonctionnment général du programme. On pouvait penser que cet apprentissage en vue de se familiariser avec l'appareil et le logiciel aurait pu être pour le sujet, une source d'ambiguité dans la consigne. Le tableau 6.12 démontre qu'il n'en est rien et qu'au contraire, il semblerait que les directives de la version informatisée soient plus claires. De fait, en assignant une valeur numérique aux catégories proposées (1 pour «Très claires» et 3 pour «Ambigües»), on trouve une moyenne de 1.33 pour la version conventionnelle et de 1.22 pour la version informatisée. Toutefois, le test *t* indique que la différence ne saurait être significative (p =.263).

TABLEAU 6.12
Clarté de la consigne (#10)

	Très claires	Généralement claires	Ambigües	Total
3.1	72	3	13	88
STRAT	78	1	9	88
Total	150	4	22	176

X^2 = 1.967 dl = 2 p = .374

En faisant des recoupements avec l'âge, on observe que les sujets les plus âgés (plus de 25 ans) trouvent les directives de la version 3.1 plus claires que celles fournies par la machine: plus de 93% pensent que les directives de la version conventionnelle sont très claires et 80% seulement pensent que les directives de la version informatisée sont très claires. Chez les plus jeunes (25 ans et moins) les pourcentages sont inversés: 80% disent que les directives de la version conventionnelle sont très claires alors que le taux dépasse 90% en ce qui a trait aux directives données par la machine. Il convient également de signaler que chez les sujets dont l'anglais

n'est pas la langue maternelle, 93% estiment que la machine fournit des directives très claires et aucun ne qualifie ces directives d'ambigües. Par conséquent, même si on leur explique une consigne relativement compliquée, dans leur langue seconde, ces sujets ne sont pas défavorisés lorsqu'on leur soumet le test informatisé.

— **Question #11:** *Vous sentiez vous à l'aise dans l'environnement du test (classe, feuille de réponses, surveillant... vs salle d'informatique, clavier, écran...)? Très détendu, détendu, tendu ou très tendu?*

Il est raisonnable de penser qu'en situation de test l'étudiant sera tendu et qu'il le sera d'autant plus que cette situation présente des éléments inconnus ou inattendus. On s'attendrait donc à ce que les étudiants qui font la version adaptative montrent un niveau de stress élevé. Or, il n'en est rien. D'une part, personne ne se déclare «Très tendu», peu importe le test. Il est certain qu'un test de classement génère moins d'anxiété que d'autres types de tests d'autant plus que dans la situation de l'expérimentation, ce test ne risquait pas d'avoir des effets importants sur les notes des étudiants. D'autre part, en faisant les moyennes des valeurs numériques attribuées à chaque catégorie (où 1 correspond à «Très détendu» et 4 à «Très tendu», on constate que la moyenne du test conventionnel (1.68) est supérieure à celle du test informatisé (1.59). De fait, si la différence n'est pas significative, c'est en partie parce le nombre des sujets est limité et que la variance des réponses pour la version informatisée est plus grande. Le tableau 6.13 montre d'ailleurs une certaine polarisation des positions des réactions face à la version STRAT. Si on en juge d'après la valeur du chi-carré, cette polarisation risque peu d'être le simple fruit du hasard. On voit donc que non seulement la situation du test adaptatif ne cause pas plus d'anxiété, mais que dans certaines circonstances, elle pourrait même la réduire.

TABLEAU 6.13
Niveau d'anxiété (#11)

	Très détendu	Détendu	Tendu	Très tendu	Total
3.1	31	54	3	0	88
STRAT	42	40	6	0	88
Total	73	94	9	0	176

X^2 = 4.743 dl = 2 p = .093

En prenant en considération les variables socio-démographiques, on constate que 50% des sujets qui se spécialisent en français affirment être très détendus au test «papier-crayon» mais qu'un peu moins de 45% affirment être très détendus au test adaptatif. Par contre, seulement 28% des non spécialistes se croient très détendus avec la version «papier-crayon» et près de 50% avec la version adaptative. En d'autres termes, l'utilisation d'une version informatisée avec les non-spécialistes permettrait d'atteindre le niveau minimal d'anxiété que les spécialistes, plus habitués aux épreuves traditionnelles en français, atteignent avec la version «papier-crayon». Par ailleurs, la répartition des réponses en tenant compte à la fois du stress et de l'expérience avec l'ordinateur ne permet pas de conclure, comme on pourrait le croire, que les usagers réguliers des ordinateurs soient plus détendus avec le test informatisé. Enfin, il est faux de croire que les sujets les plus âgés puissent être plus anxieux avec le test informatisé. Les résultats confirmeraient plutôt la tendance inverse. En effet, plus de la moitié des sujets âgés de plus de 25 ans se disent très détendus au test informatisé alors que seulement le tiers se disent aussi détendus au test conventionnel; quant au 20 ans et moins, si la même proportion se disent très détendus au test conventionnel, 42%, soit 12% de moins que les ainés, affirment être très détendus au test informatisé.

— **Question #12:** *Vous était-il difficile de vous concentrer pendant le test? Très difficile, difficile, facile ou très facile?*

Il nous semblait important de questionner les étudiants sur cet aspect puisque l'environnement peut avoir des effets importants sur la capacité de concentration et conséquemment sur les résultats au test. Le tableau 6.14 montre pourtant que le degré de concentration ne varie pas du tout d'une version à l'autre. Dans l'ensemble, les sujets n'éprouvent pas de problèmes de concentration. D'autre part, la cause des problèmes de ceux qui en éprouvent ne réside sûrement pas dans le mode d'administration.

TABLEAU 6.14
Capacité de concentration (#12)

	Très difficile	Difficile	Facile	Très facile	Total
3.1	1	17	54	16	88
STRAT	1	19	52	16	88
Total	2	36	106	32	176

$X^2 = 0.149$ $dl = 3$ $p = .985$

L'examen des interactions avec les autres variables n'est guère plus révélateur: aucun sous-groupe ne semble avoir connu de problème de concentration particulier.

– **Question 13:** *Que faisiez vous le plus souvent quand vous ignoriez la bonne réponse? Deviner en choisissant a, b, c ou d au hasard, deviner en répétant toujours la même réponse ou omettre la réponse?*

Il est permis de croire que le fait d'utiliser un médium différent peut affecter les stratégies de réponse dans les cas où l'étudiant ne connaît pas la réponse. Par exemple, il semble plus facile d'omettre une réponse avec une feuille de réponses puisque l'étudiant n'a aucun geste à poser alors qu'avec le clavier il doit appuyer sur la barre d'espacement; de plus, dans la consigne même du test adaptatif, on invitait le sujet à utiliser cette touche le moins

possible. Par ailleurs, il peut sembler plus facile de répéter la réponse précédente avec l'ordinateur puisqu'il suffit de toujours appuyer la même touche. Il faut pourtant reconnaître que ces hypothèses ne se vérifient pas du tout avec les données que nous reproduisons dans le tableau 6.15.

TABLEAU 6.15
Stratégies de réponses (#13)

	Au hasard	Même réponse	Omission	Total
3.1	63	2	5	70
STRAT	60	2	7	69
Total	123	4	12	139

$X^2 = 0.399$ $dl = 2$ $p = .819$

Tant pour la version informatisée que pour la version conventionnelle, répondre au hasard est nettement la stratégie la plus populaire quand l'étudiant ignore la réponse. La popularité de ces sélections purement aléatoires interdit d'ailleurs tout recoupement avec les autres variables car, peu importe le sous-groupe retenu, l'emploi des deux autres stratégies reste exceptionnel. Il faut aussi noter que plusieurs sujets, n'ont pas répondu à cette dernière question, ayant peut-être l'impression de ne jamais avoir dû faire face à la situation ou ayant recours à une stratégie personnelle, différente de celles parmi lesquelles on leur demandait de choisir.

— **Commentaires:** Dans plusieurs cas, la section où l'on sollicitaient des commentaires a permis aux répondants de préciser leurs réponses à la dernière question à propos des stratégies qu'ils utilisaient quand il ne connaissaient pas la bonne réponse. Ainsi, beaucoup d'étudiants se défendaient de «deviner» précisant qu'ils procédaient en évaluant la vraisemblance de chaque distracteur (inférence et élimination).

Le tableau 6.16 résume les commentaires des étudiants. Nous avons éliminé de ce sommaire les opinions qui n'étaient exprimées qu'une fois. De plus, comme la majorité des étudiants n'ajoutaient pas de commentaire, il est difficile de tirer des conclusions claires surtout pour comparer les deux versions. Les commentaires relatifs au contenu s'appliquent à l'une ou l'autre des versions. Il est certain toutefois que la lacune que soulignent beaucoup d'étudiants quant à l'absence de la langue parlée est réelle et tient en grande partie aux limites qu'impose le test informatisé dans l'élaboration de versions comparables. Il est intéressant de noter que du point de vue de la perception du test, les deux versions sont jugées favorablement. Du point de vue de la formule, des sujets ont déploré le fait de ne pas pouvoir corriger ou réviser leurs réponses avec la version informatisée. Enfin les problèmes techniques dont deux étudiants font état (emballement du programme et panne d'un appareil) sont typiques de ce qui peut survenir pendant l'administration d'un test informatisé.

TABLEAU 6.16
Sommaire des commentaires au questionnaire

Description	Fréquence	
	3.1	STRAT
CONTENU		
Il manque une partie d'expression orale.	3	6
Il manque une partie d'écoute.	2	1
On évalue seulement la compétence à l'écrit.	2	3
Il n'y a pas assez de grammaire.	3	-
Le vocabulaire est trop complexe.	1	3
Aspects extra-linguistiques au sous-test #2.	-	2
Le test couvre tous les aspects.	1	1
PERCEPTION		
Il est difficile de juger d'un test.	-	3
Le test est intéresuant.	1	1
Le test est juste et précis.	3	5
FORMAT		
L'effet du hasard est trop important.	3	-
On ne peut pas corriger une réponse.	-	2
On ne peut pas réviser.	-	2
STRATEGIES		
Inférence ("Educated guess").	6	7
Elimination des réponses peu plausibles.	2	3
Relecture.	-	2
ENVIRONNEMENT		
Fatigue au moment de l'administration.	1	3
Problèmes techniques.	-	2

6.2.2.2.3 Conclusions de l'analyse

Il convient de rappeler que nous ne cherchions pas à détecter avec ce questionnaire, des différences marginales dans les réactions des étudiants. Nous voulions plutôt vérifier s'il y avait des raisons d'ordre psychologique évidentes qui empêchaient toute tentative de comparer les résultats obtenus avec une version «papier-crayon» et une version adaptative. Or, il semble bien que si les modes d'administration sont différents, il n'y ait pas de différences majeures dans les réactions des étudiants. Par exemple, nous n'avons pas pu confirmer que la version informatisée pouvait générer plus d'anxiété, qu'elle était jugée plus positivement ou qu'elle favorisait le développement de stratégies de réponses particulières.

Deux aspects qui pourraient faire l'objet d'une étude plus poussée ressortent pourtant. Premièrement, il semble que les étudiants qui ne se spécialisent pas en français soient plus détendus avec la version adaptative que les spécialistes. Cette tendance mériterait d'être étudiée de façon plus approfondie car elle pourrait impliquer que le passage à un test adaptatif mènerait vers une évaluation moins biaisée. En effet, les étudiants spécialistes, de par leur apprentissage formel de la langue, ont l'habitude des tests de langue écrits traditionnels et peuvent ainsi être favorisés par une version «papier-crayon».

Deuxièmement, nous avons relevé que les étudiants sont peu sensibles au fait que la version adaptative sélectionne les items en fonction de leur habileté. On peut penser que le niveau où chaque item apporte le maximum d'information ne correspond pas nécessairement au niveau où cet item est jugé le plus approprié. En d'autres termes, la difficulté relative, telle que perçue par les étudiants, ne coïncide pas avec la difficulté (présentée à la section 2.2.1) telle que mesurée lors de la calibration.

6.2.3 ***Analyse qualitative***

6.2.3.1 *L'approche qualitative*

Selon Faerch et Kasper (1987), on peut voir les productions des apprenants en langue seconde, comme une succession à

plusieurs niveaux d'une série de «produits». Dans cette perspective, on peut prétendre que les données recueillies à l'aide d'un questionnaire comme celui que nous avons utilisé ne peuvent pas vraiment rendre compte de la profondeur des processus impliqués lorsqu'un étudiant répond à une question. Cohen (1987) souligne que les rapports verbaux peuvent fournir des indications précieuses, sinon sur les processus reliés aux stratégies d'acquisition, du moins sur les processus reliés aux stratégies d'apprentissage, par nature plus conscientes que les stratégies d'acquisition. Ericsson et Simon (1987), deux psychologues pionniers de ces techniques d'observation, distinguent les techniques introspectives des techniques rétrospectives. En introspection, le sujet fait des commentaires «sur le champ», c'est-à-dire pendant qu'il accomplit la tâche (exercice de compréhension, test, rédaction...). En rétrospection, le(s) sujet(s), avec l'aide discrète d'un animateur, essaie(nt) de retracer «après-coup» la démarche intellectuelle suivie.

Cohen (1984) montre que les rapports verbaux, introspectifs ou rétrospectifs, servent à préciser ce que mesure effectivement un test. Il rapporte plusieurs études où l'on demandait aux sujets de commenter leur démarche pendant qu'ils complétaient un test de closure ou pendant qu'ils répondaient à des questions à choix multiple. Grotjahn (1987) et Feldman et Stemmer (1987) ont utilisé avec succès une approche semblable pour examiner la validité des tests «C», une forme de test lacunaire apparentée au test de closure. Quant à nous, nous avons opté pour une technique rétrospective relativement simple, la discussion de groupe. Il nous semblait que les rapports verbaux ainsi recueillis serviraient à compléter les données du questionnaire et en corrigeraient même les lacunes.

6.2.3.2 *Les résultats*

Deux groupes d'une dizaine d'étudiants anglophones inscrits dans un programme intensif de l'Université du Québec de Trois-Rivières ont fait chacun une version du test (la version 3.1 et la version STRAT). Ces étudiants se retrouvaient à tous les niveaux; cependant comme l'expérimentation se déroulait pendant la

troisième semaine du programme, il n'y avait pas de débutant absolu. Immédiatement après l'épreuve, on amorçait une discussion de groupe sur le test lui-même. Il faut noter qu'à ce moment, la machine avait déjà informé les sujets du test adaptatif de leur niveau alors que la correction du test conventionnel s'est faite après la discussion. Étant donné les règles établies par le programme, la discussion s'est déroulée en français. Il est possible que cette décision ait pu à l'occasion brimer les moins avancés, mais nous nous sommes ainsi assuré de l'entière collaboration de l'organisation du programme et des étudiants pour qui cette discussion s'intégrait dans les activités du programme.

Conformément aux protocoles suggérés pour ce type de discussions, il s'agissait d'une entrevue non interventioniste. Le rôle de l'animateur consistait essentiellement à ramener la discussion vers les aspects importants sans diriger la discussion ni même chercher à couvrir tous les aspects. Les principaux aspects que nous comptions toucher au cours de la discussion étaient les suivants:

difficulté de l'ensemble du test;
difficulté relative des items;
clarté de la consigne;
stratégies de réponse;
nervosité;
intérêt à l'égard du test;
préférence *a priori*;

Les deux discussions ont été enregistrées sur cassette, puis transcrite. Nous ne rapportons dans les lignes suivantes que les éléments qui se dégagent de l'analyse que nous avons faite du contenu des transcriptions.

6.2.3.2.1 Le test adaptatif

On a demandé à 10 étudiants de se rendre au laboratoire de micro-informatique pour faire le test STRAT. La discussion s'est amorcée immédiatement après le test, au laboratoire même. De façon générale, les étudiants semblaient avoir porté beaucoup d'intérêt au test même si une participante le trouvait «impersonnel».

Unanimement, les étudiants ont d'abord reconnu qu'ils n'avaient eu aucun problème avec la consigne, peu importe leur degré de familiarisation avec la machine. De fait, les explications que fournit le programme ont rassuré les plus réticents de sorte que personne n'a dit avoir éprouvé de crainte ou d'anxiété au cours du test. Les participants ont souligné que la nervosité qui est associée habituellement à un test tient à la signification que prend le test pour le dossier scolaire et non au mode d'administration.

Contrairement à ce qu'on pourrait attendre d'un test adaptatif, les participants ont indiqué que le niveau de difficulté de chaque item variait. Certaines questions leur paraissaient nettement plus faciles ou plus difficiles que d'autres. Tous s'entendaient pour dire que la première partie (sous-test de compréhension) était la plus difficile à cause de la complexité du vocabulaire. Les étudiants ont comparé le test avec le *test Laval* qui leur avait été administré comme test de classement au début du programme. Selon les plus débutants, le *test Laval* était beaucoup plus difficile alors que selon les plus avancés, le *test Laval* était plus facile. On reconnaît ici l'effet de la procédure adaptative du point de vue de la perception de la difficulté générale du test. Interrogés quant à la proportion de questions auxquelles ils pensaient avoir répondu correctement, les étudiant s'entendaient pour dire que le taux se situait autour de 80%. Une seule personne a remarqué que le test s'adaptait à son niveau; les autres ont manifesté une certaine surprise en apprenant qu'ils avaient tous fait des tests différents.

Tous ont apprécié de connaître leur résultat sitôt le test terminé. Il ont tous fait remarquer que le niveau qui leur avait été communiqué correspondait à ce qu'il croyait être leur niveau réel sauf pour ce qui est de la performance à l'oral. D'après l'ensemble du groupe, il faudrait mesurer la conversation car c'est la facilité de converser dans la langue seconde qui serait la principale source d'hétérogénéité des groupes-classes.

Quand on leur a demandé s'ils auraient préféré faire la version conventionnelle, les étudiants ont hésité. Plusieurs ont fait remarquer que les stratégies de test qu'ils utilisaient habituellement

ne fonctionnaient pas puisqu'ils étaient dans l'impossibilité de faire de révisions. Ils ont noté que le mode de présentation empêchait l'utilisation d'éléments des autres questions comme indices; les items seraient donc tout à fait indépendants. Par ailleurs, aucun étudiant n'aurait utilisé la barre d'espacement pour omettre une réponse. Quand ils ignoraient la réponse, disaient la plupart, ils tentaient de repérer les deux distracteurs les moins plausibles pour restreindre le choix à une simple alternative.

6.2.3.2.2 Le test conventionnel

Pendant que le groupe qui se trouvait au laboratoire de micro-informatique faisait le test adaptatif puis en discutait, un deuxième groupe de 11 étudiants travaillait avec la version 3.1. La dernière copie remise, on a entamé la discussion autour du format du test. Tous ont admis qu'avec un test à choix multiple comme celui qu'ils venaient de faire, la tâche était très claire. Cependant, on a remis en question la précision des tests à choix multiple, d'autant plus qu'il n'y avait pas de partie orale. On ne savait pas si cette partie orale devait mesurer la compréhension ou l'expression mais un concensus s'est établi quant à la nécessité d'ajouter une composante orale. Malgré cette réserve quant au contenu, tous désiraient connaître leur résultat et ont par la suite attendu que la correction soit complétée.

Personne n'était nerveux et les participants ont rappelé que dans ce contexte il y avait peu de raison de l'être parce que la nervosité est reliée à la signification du test. Une étudiante a mentionné toutefois qu'elle aurait été nerveuse si on lui avait proposé un test informatisé. La majorité des participants ont par contre déclaré que, par curiosité, ils auraient opté pour cette dernière version s'ils avaient eu le choix.

Une seule étudiante avait remarqué la progression de la difficulté des items (du plus facile au plus difficile). Tous ont affirmé que la première partie exigeait beaucoup de lecture et qu'à cause de la complexité du vocabulaire, elle était nettement plus difficile. Il est

à noter que comme pour la version informatisée, les plus débutants ont trouvé ce test plus difficile que le *test Laval* tandis que les plus avancés le trouvaient plus facile. Cette observation nous amène à penser que la perception de la difficulté générale du test par rapport au test de référence que représente le *test Laval*, tient au contenu plutôt qu'au mode de sélection des items. Par contre, interrogés sur la proportion des items qu'ils croyaient avoir réussis, les participants ont donné des réponses très variées: trois d'entre eux estimaient leur taux de réussite à 50%, tandis qu'une participante pensait n'avoir presqu'aucune réponse correcte et qu'une autre pensait avait trouvé presque toutes les réponses.

En ce qui a trait aux stratégies mises en oeuvre, la technique d'élimination des deux distracteurs les moins plausibles a de nouveau été mentionnée. On a nuancé toutefois en ajoutant qu'on cherchait parfois des mots clés et des indices, y compris des indices fournis par d'autres items. Par ailleurs, quoique conscients de la possibilité de faire des révisions ou des retours, tous les participants, à l'exception d'un seul (de niveau avancé), ont avoué ne pas avoir exploité cette possibilité.

Enfin, les participants n'ont pas trouvé le test trop long. Il ont signalé que la deuxième partie, malgré les jugements extra-linguistiques qu'elle impliquait, leur avait paru la plus intéressante.

6.2.3.2.3 Conclusions de l'analyse

Aux yeux de l'observateur, les sujets paraissaient plus critiques envers le format à choix multiple avec la version conventionnelle qu'avec la version informatisée pour laquelle il imaginaient peut-être mal une autre formule. Il est clair qu'aucun des tests ne suscitait de nervosité ou d'anxiété et que la version informatisée éveillait davantage l'intérêt des sujets. Par ailleurs, les commentaires quant au contenu ne changeaient guère. Tous étaient d'accord avec la nécessité d'ajouter une composante orale. Tous s'entendaient également sur le fait que la première partie était nettement plus difficile.

Ce dernier commentaire sur la difficulté du premier sous-test peut surprendre chez ceux qui ont fait la version adaptative. En effet, en principe, les items des trois sous-tests se réfèrent à une échelle commune et c'est le niveau du sujet qui détermine la difficulté des items. On voit donc que la difficulté subjective des items ne correspond pas à leur difficulté objective (le paramètre *b*). Mais comment expliquer que, d'une part, les débutants et les avancés comparent les deux versions de la même façon par rapport au *test Laval* mais que, d'autre part, ils font des prévisions assez justes quant à la proportion des items qu'ils ont réussis? Une hypothèse de recherche s'ouvre: la perception de la difficulté d'un test (et peut-être de toute tâche langagière) ne dépendrait pas tant de considérations probabilistes (les chances de réussir) que de considérations touchant la nature même de la tâche. Dans cette perspective, concevoir l'adaptabilité d'un test à partir uniquement des paramètres de chaque item ne suffit plus pour en arriver à un test «sur mesure».

Enfin, il est intéressant de noter qu'on reproche au test informatisé d'empêcher les retours et les révisions puisqu'une fois qu'on a appuyé sur une touche, l'item est irrécupérable. Même si beaucoup d'étudiants disent ne pas avoir fait de retours ou de révisions de façon systématique, ils signalent que la recherche d'indices à travers tout le test fait partie des stratégies dont ils disposent. Ainsi, bien qu'on puisse prétendre satisfaire l'exigence d'indépendance des items avec la version informatisée, on ne saurait en dire autant de la version conventionnelle. Cela devient problématique quand on sait que la calibration s'est effectuée à partir d'un test «papier-crayon». Enfin, en commentant leurs stratégies de réponse, les sujets sont unanimes pour dire qu'il n'ont jamais répondu tout à fait au hasard mais qu'ils préfèrent procéder par inférence (*educated guess*). Que représente donc réellement le paramètre c dans une analyse selon un modèle à trois paramètres si personne ne devine vraiment? Quel sont les conséquences de l'emploi d'un tel paramètre si le hasard ne joue absolument pas? Voilà des considérations qui remettent en question l'applicabilité de la théorie du trait latent dans l'élaboration d'un test adaptatif en langue seconde.

Une approche qualitative paraît particulièrement appropriée pour examiner la question de la difficulté subjective et l'emploi des stratégies d'inférence. On pourrait ainsi, dans une étude ultérieure, recourir à une méthode introspective: pendant que les sujets feraient le test, on leur demanderait de commenter leur démarche intellectuelle à haute voix.

6.3 **Le plan administratif**

Nous n'avons pas fait d'expérimentation avec comme objectif spécifique de comparer les deux versions du point de vue administratif. Toutefois, les expérimentations que nous avons faites dans le milieu, dans un premier temps pour mettre au point les instruments et dans un deuxième temps pour en faire une étude comparative, nous ont amené à prendre contact avec les milieux où ces tests pourraient être utilisés. De ce point de vue, par expérience plutôt que par expérimentation, nous avons amassé un ensemble de données nous permettant de tirer certaines conclusions quant aux conditions d'utilisation éventuelles des instruments que nous avons mis au point. Ces considérations pratiques nous semblent primordiales et bien qu'elles n'aient pas fait l'objet central de notre recherche, nous ne devons pas les ignorer.

6.3.1. *Le déroulement de l'expérimentation*

La phase initiale de notre projet consistait à mettre sur pied les instruments de mesure que nous avions l'intention de comparer. À ce moment, la théorie du trait latent nous semblait être le cadre psychométrique le plus approprié. De plus, nous en sommes venu rapidement à penser que seul le modèle à trois paramètres convenait. L'adoption d'un tel modèle pose, du moins dans le contexte de l'éducation post-secondaire au Canada, des problèmes pratiques sérieux car il est extrêmement difficile de trouver les échantillons requis pour la calibration des items. Réunir un millier de sujets présentant des caractéristiques communes, est un objectif difficilement réalisable de sorte que nous avons dû nous contenter d'un échantillon plus modeste d'environ 750 sujets. Quant à l'addition

éventuelle de nouveaux items, si elle est éminemment souhaitable, elle suppose la mise en place d'un schéma d'ancrage qui requiert la collaboration d'un si grand nombre de sujets qu'il faudra s'accommoder d'un échéancier très long. Chercher à accélérer le processus risque de poser des problèmes éthiques importants du fait qu'on doit soumettre la population étudiante à un exercice dont elle profite peu.

Il nous semble que la version 2 ne pouvait pas être plus longue. Il est impensable qu'un même étudiant réponde à plus de 150 items. La durée des versions 3.1 et 3.2 nous semble mieux correspondre à ce qu'on peut raisonnablement attendre d'un test de classement. Les étudiants mettent généralement moins d'une heure pour faire l'épreuve. Néanmoins, afin de ne pas imposer une contrainte de temps qui risque de fausser les résultats, il est sage de prévoir une période d'une heure et demie pour l'administration des versions 3.1 ou 3.2. La version adaptative présente, de ce point de vue, un avantage considérable. Lors de notre expérimentation à Saint-Georges nous avions prévu des changements de groupe à toutes les demi-heures. Cet horaire s'est avéré assez réaliste bien qu'il eût été souhaitable de réserver au moins un poste de travail pour les étudiants plus lents. Toutefois, dans l'ensemble, on peut dire que la version informatisée demande deux fois moins de temps.

La surveillance des tests «papier-crayon» est on ne peut plus simple. Il suffit de distribuer le matériel du test et de s'assurer de tout récupérer à la fin. Hormis une rectification pour une erreur dans le premier exemple, le surveillant n'a jamais eu à intervenir. Il faut par la suite prévoir du temps pour la correction. Nous avons préparé des acétates qui permettent de corriger rapidement les copies. Cependant nous avons pu constater que ce mode de correction n'est pas infaillible. Avec certains groupes, nous avons trouvé jusqu'à 10% des scores calculés par cette correction manuelle qui étaient erronés et qu'on a dû recorriger pour le traitement des données. À cet égard, la version informatisée est nettement supérieure parce la correction est immédiate, économique et sans erreur. Par contre, nous nous sommes rendu compte que l'administration du test informatisé en groupe n'est pas aussi automatisée qu'on pourrait le croire. À quelques reprises, le surveillant a dû intervenir

pour aider des étudiants qui éprouvaient des problèmes. Dans certains cas, des étudiants, habitués à d'autres logiciels, étaient déroutés par le fait que la touche <Return> soit inopérante. Dans d'autres cas, plus sérieux, les étudiants maintenaient le doigt appuyé sur une touche provoquant ainsi l'emballement du programme. Enfin, bien qu'on ait cherché dans la programmation à minimiser l'utilisation du clavier, l'étudiant doit inévitablement s'identifier; or, certains ont eu de la difficulté à taper leur nom. Cela met en évidence une lacune fondamentale de la technologie actuelle: tant que le clavier restera le principal moyen par lequel on transmet l'information à la machine, certains étudiants seront défavorisés et les applications pédagogiques de l'ordinateur seront fort limitées.

Enfin, il faut noter que lors des deux expérimentations menées dans un laboratoire de micro-informatique, un appareil a fait défaut alors qu'un étudiant y était installé. De pareilles défaillances sont des aléas avec lesquels il faut composer. Ainsi compte tenu à la fois des défaillances qui peuvent survenir et des problèmes que peuvent connaître les étudiants, il est difficile d'imaginer une administration en groupe sans surveillance. En utilisant un seul appareil, on peut espérer se dispenser de surveillance mais il reste qu'une personne devrait être disponible au cas où l'étudiant aurait besoin d'aide.

6.3.2 *Les ressources et les besoins*

L'administration individuelle est sans doute la formule la plus attrayante pour les établissements intéressés au testing adaptatif. En effet, peu de programmes de langue seconde peuvent compter, parmi leurs ressources, l'accès facile à un laboratoire de micro-informatique. Le programme *CAPT* aurait donc sa place dans les programmes où les tests de classement sont administrés individuellement à différents moments. Par contre, lorsque qu'un grand nombre d'étudiants doivent être triés, l'administration en groupe d'une version «papier-crayon» est nettement plus avantageuse.

Par ailleurs, il faut se demander si ce test, tel que conçu pour satisfaire les exigences de la présente recherche, répond aux besoins des établissements post-secondaires. On peut distinguer deux orientations tout à fait opposées parmi les programmes qui s'adressent à la population de niveau post-secondaire. D'une part, un certain nombre s'inspirent des approches traditionnelles et risquent de ne pas trouver leur compte dans un test qui ne mesure pas spécifiquement les éléments grammaticaux. On remarque de toute façon que ces programmes tendent à regrouper les étudiants selon une séquence de cours pré-déterminée ou en fonction du nombre d'années d'étude de la langue. D'autre part, des programmes mettent l'accent sur la compétence à l'oral et pourraient reprocher aux instruments que nous avons élaborés de ne pas comporter de composante orale. Bien que ces instruments mesurent la maîtrise générale, il est souhaitable, dans la perpective d'un enseignement de l'oral, de pouvoir aussi mesurer plus spécifiquement la capacité de comprendre et de s'exprimer oralement. Ainsi, si le test devait être utilisé par ces établissements, il faudrait y ajouter un sous-test de compréhension auditive et/ou le compléter par une entrevue.

Enfin, il ne faut pas oublier qu'après avoir utilisé régulièrement les deux mêmes formes d'un test «papier-crayon» ou un test adaptatif construit à partir d'une banque comprenant peu d'items, on devra songer à créer des formes parallèles supplémentaires ou à élargir la banque. Étant donné les coûts associés à une telle entreprise, il n'est pas certain que les programmes de langue seconde, traditionnellement considérés comme les parents pauvres des établissements post-secondaires, puissent financer l'élaboration de tests semblables à ceux que nous avons mis au point. En ce sens, le testing adaptatif, comme toute autre approche exploitant une banque d'items pré-calibrés, ne se prête pas au testing sur une échelle réduite.

CONCLUSION

La présente recherche s'est déroulée en trois phases. Dans un premier temps, il nous a fallu préciser notre cadre théorique. D'une part, nous estimions qu'un test de classement devait évaluer une maîtrise générale de la langue et qu'un tel attribut était effectivement mesurable. D'autre part, compte tenu des recherches actuelles dans le domaine docimologique, la théorie du trait latent nous paraissait le mieux convenir à notre objectif. Dans un deuxième temps, il nous a fallu mettre au point les instruments. Nous avons élaboré une version expérimentale qui a par la suite servi à créer deux versions conventionnelles parallèles. Les items retenus ont également été intégrés dans les deux versions adaptatives que nous avons par la suite programmées.

Ce n'est que dans un troisième temps que nous avons pu comparer les deux types de tests: avec et sans ordinateur. Du point de vue théorique, il nous semblait que l'emploi d'un test informatisé adaptatif permettrait d'obtenir des tests plus courts mais aussi précis, mieux adaptés au niveau de chaque étudiant et simples à administrer. Par contre, nous avions des réserves quant au contenu et à la nature des items. De plus, nous craignions des effets psychologiques négatifs et nous nous interrogions sur l'aspect pratique. À l'aide de données expérimentales, nous avons pu confirmer que le test adaptatif par stratification était généralement tout aussi fiable tout en étant deux fois plus court. Cependant, il ne nous apparaissait pas possible de comparer les résultats obtenus avec des procédures d'administration différentes car les modes de correction ne sont pas identiques. Au plan psychologique, nous avons observé que l'ordinateur ne produisait pas les effets négatifs que nous anticipions. Il nous est aussi apparu que la notion de difficulté sur laquelle repose le testing adaptatif, doit être revue parce que la perception de la difficulté ferait intervenir des juge-

ments globaux qui dépassent les caractéristiques statistiques individuelles des items. Les résultats de l'expérimentation révèlent la nature exploratoire de la présente recherche puisque certains aspects devront être approfondis et pourraient faire l'objet d'études ultérieures.

Considérant les forces et les faiblesses des deux types d'administration que nous avons comparés, il nous semble que le test adaptatif peut être une alternative intéressante dans l'optique de l'administration individuelle d'un test de classement. Une version conventionnelle est certainement plus appropriée pour des administrations en groupe. De plus, nous doutons qu'une procédure adaptative puisse être appliquée dans le cadre d'un test diagnostique où la division du contenu implique la multidimensionalité. Nous doutons également qu'une procédure adaptative puisse servir dans le cadre d'un test de certification où il est essentiel de viser l'authenticité des tâches soumises au candidat. Enfin, étant donné le contenu du test, celui-ci risque de ne pas suffire pour mener à une décision juste. Par exemple, s'il doit servir à classer un étudiant à l'intérieur d'une séquence de cours de conversation, il serait souhaitable que le test soit complété par une entrevue.

Que les instruments que nous avons élaborés mesurent effectivement la maîtrise générale sans pouvoir toujours mener à une décision de classement juste s'explique certes par la multiplicité des approches et des programmes d'enseignement. Mais cela est aussi attribuable aux compromis que nous avons dû faire pour les fins de l'expérimentation. Afin de comparer les procédures d'administration, nous avons conçu des items qui pouvaient être aisément transposés à l'ordinateur. Cette situation traduit, nous semble-t-il le paradoxe du testing adaptatif. D'une part, l'utilisation de l'ordinateur offre de grandes possibilités: graphiques, animation, couleur... Voilà, pour reprendre l'expression de Canale (1985) la «promesse» du testing adaptatif. D'autre part, les exigences de la théorie du trait latent sont telles qu'on doit souvent renoncer à exploiter ces possibilités. Par exemple, des considérations pratiques nous forcent généralement à calibrer à partir d'une version «papier-crayon». Plus encore, la règle d'unidimensionalité impose à cette version «papier-crayon»

un cadre plus étroit que celui qui régit les tests conventionnels. C'est l'envers de la médaille, la «menace» du testing adaptatif selon Canale. Il faut cependant espérer que des travaux en docimologie tels que ceux de Wilcox (1981) de Reckcase (1983) ou de Traub et Lam (1985) permettront d'assouplir ou de déborder le cadre de la théorie du trait latent.

Faisant le point sur l'utilisation de l'ordinateur en évaluation de la langue seconde, Alderson (1988) commet une omission importante en effleurant à peine le concept de testing adaptatif. Il n'en reste pas moins que, comme lui, nous croyons qu'il faut transgresser les limites de cette approche pour innover dans le domaine et proposer de nouvelles formes d'activités évaluatives qui exploitent les possibilités de la machine et intègrent les nouvelles approches de la didactique des langues secondes. Enfin, il ne faut pas oublier que si les progrès technologiques en informatique sont fascinants, l'ordinateur reste dans le domaine de l'enseignement des langues un outil relativement rudimentaire qui ne doit pas nous imposer ses limites. Comme le signale Churchill (1986:20): «À force de vouloir appliquer l'ordinateur partout, on fait ce qu'on peut plutôt que ce que l'on devrait».

BIBLIOGRAPHIE

Adams, M.L. Five cooccuring factors in speaking proficiency. In J.R. Firth (réd.), *Measuring Spoken Language Proficiency*. Washington, DC: Georgetown University Press, 1980, 1-6.

Alderson, C.J. Native and non-native speaker performance on Cloze tests. *Language Learning*, 1980, *30*, 59-76.

_____ The Cloze procedure and proficiency in English as a foreign language. In J.W. Jr Oller (réd.), *Issues in Language Testing Research*. Rowley, MA: Newbury House, 1983, 205-217.

_____ *Innovation in Language Testing: Can the Microcomputer help?* (Special report No.1, Language testing update). Lancaster, UK: University of Lancaster, 1988.

_____ *Judgements in Language Testing*. Communication au Congrès de l'AILA, Thessalonique, Grèce, avril 1990.

Alderson, J.C. et Urquhart, A.H. The effect of students' background discipline on comprehension: A pilot study. In A. Hughes et D. Porter (réd.), *Current Developments in Language Testing*. Londres: Academic Press, 1983, 121-127.

_____ This test is unfair: I'm not an economist. In P. Hauptman, R. Leblanc et M. Wesche (réd.), *L'évaluation de la performance en langue seconde*. Ottawa, ON: Éditions de l'Université d'Ottawa, 1985, 25-43.

Allen, M.J. et Yen, W.M. *Introduction to Measurement Theory*. Monterey, CA: Brook/Cole, 1979.

Amarel, M. The classroom: an instructional setting for teachers, and the computer. In C.A. Wilkinson (réd.), *Classroom Computers and Cognitive Science*. New York: Academic Press, 1983, 15-28.

Anastasi, A. *Psychological Testing* (2nd ed.). New York: MacMillan, 1961.

Anderson, J.R. *Cognitive Psychology and its Implications*. New York: W.H. Freeman and Co., 1985.

Angoff, W.H. Scales, norms, and equivalent scores. In R.L. Thorndike (réd.), *Educational Measurement* (2ème éd.). Washington, DC: American Council on Education, 1971, 508-600.

_____ Summary and derivation of equating methods used at ETS. In P. Holland et D.B. Rubin (réd.), *Test Equating*. New York: Academic Press, 1982, 55-69.

Angoff, W.H. et Sharon, A.T. A comparison of scores earned on the Test of English as a foreign language by native American college students and foreign applicants. *TESOL Quarterly*, 1971, 5, 129-136.

Assessment System Corp. *User's Manual for the MicroCAT Testing System* (2nd ed.). St-Paul, MN, 1987.

Auger, R. et Séguin, S.S. Le modèle de Rasch et la paramétrisation d'une banque d'items et d'instruments de mesure. *Mesure et évaluation en éducation*, 1986, *9*, 2/3:59-97.

Bachman, L.F. The trait structure of Cloze tests scores. *TESOL Quarterly*, 1982, *16*, 61-70.

_____ *Fundamental Considerations in Language Testing*. New York: Oxford University Press, 1990.

Bachman, L.F. et Palmer, A.S. The construct validation of the FSI oral interview. *Language Learning*, 1981, *31*, 67-86.

_____ The construct validation of some component of communicative proficiency. *TESOL Quarterly*, 1982, *16*, 449-465.

_____ Some comments on the terminology of language testing. In C. Rivera (réd.), *Communicative Competence Approaches to Language Proficiency Assessment: Research and Application*. Clevedon: Multilingual Matters, 1984, 34-43.

Bachman, L.F. et Savignon, S.J. The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal*, 1986, *70*, 380-390.

Bailey, K.M. If I had known what I know now: Performance testing of foreign teaching assistants. A. Hughes et D. Porter (réd.), *Current Developments in Language Testing*. Londres: Academic Press, 1983, 153-180.

Baker, F.B. *The Basics of Item Response Theory*. Portsmouth, NH: Heinemann, 1985.

Barnett, M.A. Syntactic and lexical/semantic skills in foreign language reading. *Modern Language Journal*, 1986, *70*, 343-349.

Bejar, I.I. A comparison of conventional and computer-based achievement testing. In D.J. Weiss (réd.), *Proceeding of the 1977 Conference on Computerized Adaptive Testing*. Minneapolis, MN: University of Minnesota, 1978, 373-385.

_____ A procedure for investigating the unidimen sionality of achievement test based on parameter estimates. *Journal of Educational Measurement*, 1980, *17*, 283-296.

_____ Introduction to item response theory models and their assumptions. In R.K. Hambleton (réd.), *Applications of Item Response Theory*. Vancouver, BC: Vancouver Educational Research Institute of British Columbia, 1983, 1-23.

Bejar, I.I. et Weiss, D.J. *A Construct Validation of Adaptive Achievement Testing* (Research Report 78-4). Minneapolis, MN: University of Minnesota, Department of Psychology, 1978.

Bejar, I.I., Weiss, D.J. et Gialluca, K.A. *An Information Comparison of Conventional and Adaptive Tests* (Research Report 77-7). Minneapolis, MN: University of Minnesota, Department of Psychology, 1977.

Berry, K.J. et Mielke, P.W. A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement*, 1988, *48*, 921-933.

Bernier, J.J. *Théorie des tests: Principes et techniques de base* (2ème éd.). Chicoutimi, QC: Gaëtan Morin, 1985.

Betz, N.E. et Weiss, D.J. *Simulation Studies of Two-Stage Ability Testing* (Research report 74-4). Minneapolis, MN: University of Minnesota, Department of Psychology, 1974.

Binet, A. *Les idées modernes sur les enfants*. Paris: Flammarion, 1909.

Birnbaum, A. Some latent-trait models and their use in inferring an examinee's ability. In F.M. Lord et M.R. Novick, *Statistical Theories of Mental Tests Scores*. Reading, MA: Addison-Wesley, 395-549.

Biskin, B.H. et Kolotkin, R.L. Effects of computerized administration on scores on the Minnesota Multiphasic Personnality Inventory. *Applied Psychological Measurement*, 1977, *1*, 543-549.

Bock, R.D. et Mislevy, R.J. Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 1982, *6*, 431-444.

Bolus, R.E., Hinofotis, F.B. et Bailey K.M. An introduction to generalizability theory in second language research. *Language Learning*, 1982, *32*, 245-268.

Bondarock, J., Child, J.R. et Tetreault, E.W. Contextual Testing. In R.L. Jones et B. Spolsky (réd.), *Testing Language Proficiency*. Arlington, VA: Center for Applied Linguistics, 1975, 89-108.

Bonnet, A. *L'intelligence artificielle: promesses et réalités.* Paris: Inter-Éditions, 1984.

Borland International Inc. *Turbo-Pascal Database Toolbox* (1ère éd.). Scotts Valley, CA, 1985.

_____ *Turbo-Pascal, Version 4.0.* Scotts Valley, CA, 1985.

Bowen, D. The effect of environment on proficiency testing. *Workpapers in Teaching English as a Second Language* (UCLA), 1978, *12*, 57-61.

Brennan, R.L. *Elements of Generalizability Theory.* Iowa City: The American College Testing Program, 1983.

Briere, E.J. Current trends in second language testing. *TESOL Quarterly*, 1969, *3*, 333-340.

Brown, J.D. Relatives merits of four methods for scoring Cloze tests. *Modern Language Journal*, 1980, *64*, 311-317.

_____ Improving ESL placement tests using two perspectives. *TESOL Quarterly*, 1989, *23*, 65-83.

Bunderson, C.V., Inouye, D.K. et Olsen J.B. The four generations of computerized educational measurement. In R.L. Linn (réd.), *Educational Measurement* (3rd ed.). New York: American Council on Education – Macmillan, 1989, 367-408.

Byrnes, H. et Canale, M. (réd.). *Defining and Developing Proficiency: Guidelines, Implementation and Concepts.* Lincolnwood, IL: National Textbook, 1987.

Canale, M. Communication: How to evaluate it? *Bulletin de l'ACLA*, 1981, *3*, 2:77-94. (a)

_____ The method effect in communicative testing. *Médium*, 1981, *6*, 4:43-55. (b)

_____ From communicative competence to communicative language pedagogy. In J. Richards et R. Smith (réd.), *Language and Communication.* Londres: Longman, 1983, 2-28.

_____ Considerations in the testing of reading and listening proficiency. *Foreign Language Annals*, 1984, *17*, 349-357.

_____ Language assessment: The method is the message. In D. Tannen et J.E. Alatis (réd.), *Georgetown University Press Roundtable on Language and Linguistics.* Washington, DC: Georgetown University Press, 1985, 259-262.

_____ The promise and threat of computerized adaptive assessment of reading comprehension. In C.W. Stansfield (réd.), *Technology and Language Testing.* Washington, DC: TESOL, 1986, 29-46.

_____ *L'évaluation naturelle.* Communication au colloque de l'AQEFLS. Montréal, mars 1988.

Canale, M. et Swain, M. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics.* 1980, *1*, 1-47.

Carroll, B.J. *Testing Communicative Performance.* Oxford: Pergamon, 1980.

_____ Is there another way? In B. Heaton (réd.), *Language Testing.* Middlesex: Modern English Publ., 1982, 1-10.

_____ Second language performance testing for university and professional contexts. In P. Hauptman, R. Leblanc et M. Wesche (réd.), *L'évaluation de la performance en langue seconde.* Ottawa, ON: Éditions de l'Université d'Ottawa, 1985, 73-88.

Carroll, J.B. Fundamental considerations in testing for English language proficiency of foreign students. In H.B. Allen (réd.), *Teaching English as a Second Language.* McGraw-Hill, 1965, 30-40.

_____ Psychometric theory and language testing. In J.W. Jr Oller (réd.). *Issues in Language Testing Research.* Rowley, MA: Newbury House, 1983, 80-107.

_____ Psychometric theory and language testing. In R.C. Grotjahn, C. Klein-Bradley et D.K. Stevenson (réd.). *Taking their Measure: The validity and validation of language tests.* Bochum: Studienverlag Dr N. Brockmeyer, 1987, 1-40.

Cartier, F.A. Criterion-referenced testing of language skills. *TESOL Quarterly,* 1968, *2*, 18-24.

_____ Alternative methods of oral proficiency assessment. In J.R. Firth (réd.), *Measuring Spoken Language Proficiency.* Washington, DC: Georgetown University Press, 1980, 7-14.

Cazabon, B. La compétence communicative et la possibilité de mesurer la variation. *Revue de l'Université Laurentienne.* 1984, *17*, 103-112.

Cazabon, B. et Size-Cazabon, J. Comment mesure-t-on la qualité de la communication? Test à l'intention des professeurs d'immersion. *Mesure et évaluation en éducation.* 1986, *9*, 4:5-39.

Chapelle, C. et Jamieson, J. Computer-assisted language learning as a predictor of success in acquiring English as a second language. *TESOL Quarterly,* 1986, *20*, 27-46.

Chapelle, C. et Roberts, C. Ambiguity tolerance and field dependance as predictors of proficiency in English as a second language. *Language Learning,* 1986, *36*, 27-45.

Chastain, K.D. Evaluating expressive objectives. *Revue canadienne des langues vivantes*, 1977, 34, 62-70.

Chomsky, N. *Aspects of the Theory of Syntax*. Cambridge, MA: The M.I.T. Press, 1965.

Choppin, B.H. Item bank using sample-free calibration. *Nature*, 1968, 219, 870-872.

Churchill, S. L'ordinateur au service de la didactique des langues. *Les amis de Sèvres*, 1986, 2, 13-26.

Clark, J.L.D. *Foreign-language testing: Theory and practice*. Concord, MA: Heinle & Heinle, 1972.

_____ Theoretical and technical considerations in oral proficiency testing. In R.L. Jones et B. Spolsky (réd.), *Testing Language Proficiency*. Arlington, VA: Center for Applied Linguistics, 1975, 10-28.

_____ Psychometric considerations in language testing. In B. Spolsky (réd.), *Advances in Language Testing Series: 2 — Approaches to Language Testing*. Washington, DC: Georgetown University Press, 1979, 15-30.

_____ Toward a common measure of speaking proficiency. In J.R. Firth (réd.), *Measuring Spoken Language Proficiency*. Washington, DC: Georgetown University Press, 1980, 15-26.

_____ Language testing: past and current status - Directions for the future. *The Modern Language Journal*, 1983, 67, 431-443.

_____ Toward a research and development strategy for computer-assisted language learning. *CALICO Journal*. Mars 1988, 5-23.

Clark, J.L.D. et Clifford, R.T. The FSI/ILR/ACTFL proficiency scale and testing techniques. *Studies in Second Language Acquisition*, 1988, *10*, 129-147.

Cleary, T.A., Linn, R.L. et Rock, D.A. An exploratory study of programmed tests. *Educational and Psychological Measurement*, 1969, *28*, 345-360.

Cliff, N. Evaluating Guttman scales: Some old and new thoughts. In H. Wainer et S. Messick (réd.), *Principles of Modern Psychological Measurement*. Hillsdale, NJ: Laurence-Erlbaum, 1983, 283-301.

Cliff, N., Cudeck, R. et McCormick, D. An empirical evaluation of implied orders as a basis for tailored testing. D.J. Weiss (réd.), *Proceeding of the 1977 Conference on Computerized Adaptive Testing*. Minneapolis, MN: University of Minnesota, 1978, 47-61.

Clifford, R.T. Foreign Service Institute: Factor scores and global ratings. In J.R. Firth (réd.), *Measuring Spoken Language Proficiency*. Washington, DC: Georgetown University Press, 1980. 27-30.

_____ *Oral Proficiency Profiles and Global Rating*. Communication à la Pre-Conference on Oral Proficiency Assessment, Washington, DC, mars 1981.

Cohen, A.D. On taking language tests: What the students report. *Language Testing*, 1984, *1*, 70-81.

_____ Using verbal reports in research on language learning. In C. Faerch et G. Kasper (réd.), *Introspection in Second Language Research*. Clevedon: Multilingual Matters, 1987, 82-95.

Cole, G. Do Cloze tests measure language ability? *Médium*, 1981, *6*, 4:71-82.

Compain, J, Duquette, L. et Laurier, M. *Le vidéo et le logiciel, outils d'autoperfectionnement pour les professeurs-es de langue*. Communication au colloque du CIPTE, Québec, octobre 1989.

Condon, E. The cultural context of language testing. In L. Palmer et B. Spolsky (réd.), *Papers on Language Testing 1967-1974*. Washington, DC: TESOL, 1975, 204-217.

Connors, K. Performance measure in L2: Classification and correlations. *Bulletin de l'ACLA*, 1983, *5*, 2:117-141.

Connors, K. et Toker, M.B. Analyses quantitatives des tests cloze: syntaxe et sémantique. *Revue canadienne des langues vivantes*, 1984, *40*, 245-263.

Cook, L.L., Dorans, N.J. et Eignor, D.R. An assessment of the dimensionality of three SAT-verbal test editions. *Journal of Educational Statistics*, 1988, *13*, 19-43.

Cook, L.L. et Eignor, D.R. Practical considerations regard ing the use of item response theory to equate test. In R.K. Hambleton (réd.), *Applications of Item Response Theory*. Vancouver, BC: Vancouver Educational Research Institute of British Columbia, 1983, 175-195.

Cory, C.H. et Rimland, B. Using computerized test to measure new dimensions of abilities: An exploratory study. *Applied Psychological Measurement*, 1977, *1*, 101-110.

Courchêne, R.J. et De Bagheera, J.I. Testing communicative competence: Problems and perspectives. *Médium*, 1981, *6*, 4:57-70.

Cowell, W.R. Item-response-theory pre-equating in the TOEFL testing program. In P.W. Holland et D.B. Rubin, *Test Equating*. New York: Academic Press, 1982, 149-161.

Cronbach, L.J. *Essential of Psychological Testing* (3ème éd.). New York: Harper and Row, 1970.

_____ Test validation. In R.L. Thorndike (réd.), *Educational Measurement* (2ème éd.). Washington, DC: American Council on Education, 1971, 443-507.

Currall, S.C. et Kirk, R.E. Predicting success in intensive foreign language courses. *Modern Language Journal*, 1986, *70*, 107-113.

Cziko, G.A. Psychometric and edumetric approaches to language testing. *Applied Linguistics*, 1981, *2*, 27-42.

_____ Cziko, G.A. Improving the psychometric, criterion-referenced, and practical qualities of integrative language tests. *TESOL Quarterly*, 1982, *16*, 367-379.

Dandonoli, P. ACTFL's current research in proficiency testing. In H. Byrnes et M. Canale (réd.). *Defining and Developing Proficiency: Guidelines, Implementation and Concepts*. Lincolnwood, IN: National Textbook, 1987, 75-96.

Dandonoli, P. et Rumizen, M. *There's More than One Way to Skin a CAT: Development of a Computer-Adaptive French Test in Reading*. Communication au symposium de CALICO, Colorado Springs, avril 1989.

Davidson, F.G. *An Exploratory Modeling Survey of the Trait Structures of Some Existing Language Test Datasets*. Thèse de doctorat non publiée, University of California at Los Angeles, 1988.

Davies, A. Two tests of speeded reading. In R.L. Jones et B. Spolsky (réd.), *Testing Language Proficiency*. Arlington, VA: Center for Applied Linguistics, 1975, 119-130.

De Jong, J.H.A.L. Item selection from pretest in mixed ability groups. In C.W. Stansfield (réd.), *Technology and Language Testing*. Washington, DC: TESOL, 1986, 91-107.

Dermer-Applebaum, S. et Taborek, E. Developing oral placement tests for community-based language programs. *TESL Canada Journal*. Novembre 1986, 207-223.

Des Brisay, M. Adding a communicative dimension to tests of linguistic competence. *Médium*, 1981, *6*, 4:107-116.

_____ *Developing an Item Bank for Institutional Use*. Communication à TESL Ontario, Toronto, ON, novembre 1988.

Divgi, D.R. Does the Rasch model really works for multiple choice items? Not if you look closely. *Journal of Educational Measurement*, 1984, *23*, 283-298.

Dorans, N.J. et Kingston, N.M. The effect of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. *Journal of Educational Measurement*. 1985. 22, 249-262.

Douesnard, M. et al. *Test de classement en français, langue seconde.* Montréal, QC: Centre éducatif et culturel. 1972.

Duke University, *CALIS: User's Guide* (version 2.20). Durham, NC: Humanities Computing Facility, 1989.

Duran, R.P. Some implications of communicative competence research for integrative proficiency testing. In C. Rivera (réd.), *Communicative Competence Approaches to Language Proficiency Assessment: Research and Applications*. Clevedon: Multilingual Matters, 1984, 44-58.

Educational Testing Services. *The ETS Oral Interview Book*. Princeton, NJ, 1978.

Emerson, P.L. Experience with computer generation and scoring of tests for a large class. *Educational and Psychological Measurement*, 1974. 34. 703-709.

Ericsson, K.A. et Simon, H.A. Verbal report on thinking. In C. Faerch et G. Kasper (réd.). *Introspection in Second Language Research*. Clevedon: Multilingual Matters, 1987. 24-53.

Everitt, B.S. *An Introduction to Latent Variable Models*. Londres: Chapman & Hall, 1984.

Faerch, C. et Kasper, G. The role of listening comprehension: A theoretical base. *Applied Linguistics*. 1986. 7. 257-274.

_____ From product to process – Introspective methods in second language research. In C. Faerch et G. Kasper (réd.). *Introspection in Second Language Research*. Clevedon: Multilingual Matters, 1987, 5-23.

Farhady, H. Measures of language proficiency from the learner's perspective. *TESOL Quarterly*, 1983. *16*. 43-60. (a)

_____ On the plausibilitry of the unitary language proficiency factor. In J.W. Jr Oller (réd.). *Issues in Language Testing Research*. Rowley, MA: Newbury House, 1983, 11-27. (b)

_____ New directions for ESL proficiency testing. In J.W. Jr Oller (réd.), *Issues in Language Testing Research*. Rowley, MA: Newbury House, 253-269. (c)

_____ The disjonctive fallacy between discrete-point and integrative tests. In J.W. Jr Oller (réd.), *Issues in Language Testing Research*. Rowley, MA: Newbury House. 1983, 319-322.

Feldman, U. et Stemmer, B. Think_____ aloud a_____ retrospective da_____ in C-te_____ taking: diffe_____ languages – diff_____ learners – sa_____ approaches? In C. Faerch et G. Kasper (réd.), *Introspection in Second Language Research*. Clevedon: Multilingual Matters, 1987, 251-267.

Finocchiaro, M. et Sako, S. *Foreign Language Testing: A Practical Approach*. New York: Regents, 1983.

Fishman, J.A. et Cooper, R.L. The sociolinguistics foundations of language testing. In B. Spolsky (réd.), *Advances in Language Testing Series: 2 – Approaches to Language Testing*. Arlington, VA: Center for Applied Linguistics, 1978, 31-38.

Freeman, A. *The Effect of Word-Processor Use on Written Products, Composing processes and Writing Pedagogy*. Communication au Colloque sur la pédagogie de la langue seconde, Université d'Ottawa, octobre 1988.

Friedman, C.B. *The Construct Validation of Second Language Proficiency Tests with Different Native Language Groups*. Thèse de doctorat, University Microfilm International, University of Indiana, 1984.

Gareau, C. L'évaluation de la compétence linguistique des membres des ordres professionnels au Québec. *Le français dans le monde*, 1981, *165*, 58-62.

Gendron, J.D. et al. *Test Laval: Formule A*. Québec: Presses de l'Université Laval, 1971.

Genesee, F. Psycholinguistic foundation of language assessment. In S. Seidner (réd.), *Issues in Language Assessment: Foundations and Research*. Evanston, IL: Illinois State Board of Education, 31-35.

_____ Psycholinguistic Aspect. In C. Rivera (réd.), *Communicative Competence Approaches to Language Proficiency Assessment: Research and Applications*. Clevedon: Multilingual Matters, 1984, 134-145.

Gialluca, K.A. et Weiss, D.J. *Efficiency of Adaptive Inter-Subtest Branching Strategy in Measurement of Classroom Achievement* (Research Report 79-6). Minneapolis, MN: University of Minnesota, Department of Psychology, 1979.

Girard, C., Huot, D. et Lussier-Charles, D. L'évaluation de la compétence de communication en classe de langue seconde. In C. Germain (réd.), *Études de linguistique appliquée 56*. Paris: Didier, 1984, 77-87.

Gradman, H.L. et Spolsky, B. Reduced redundancy testing: A progress report. In R.L. Jones et B. Spolsky (réd.), *Testing Language Proficiency*. Arlington, VA: Center for Applied Linguistics, 1975, 59-70.

Greaud, V.A. et Green, B.F. Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement.* 1986, *10*, 23-34.

Green, B.F. Adaptive testing by computer. In R.B. Eskstron (réd.), *New Directions for Testing and Measurement #17: Measurement. Technology and Individuality in Education.* San Francisco, CA: Jossey-Bass, 1983, 5-11. (a)

_____ The promise of tailored testing. In H. Wainer et S. Messick (réd.), *Principles of Modern Psychological Measurement.* Hillsdale, NJ: Laurence Erlbaum, 1983, 69-80. (b)

_____ Construct validity of computer-based tests. In H. Wainer et H.I. Braun (réd.), *Test Validity.* Hillsdale, NJ: Laurence Erlbaum, 1988, 77-84

Green, B.F. et al. Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 1984, *21*, 347-360.

Griffin, P.E. The use of latent trait models in the calibration of tests of spoken language in large-scale selection placement programs. In Y.P. Lee (réd.), *New Directions in Language Testing.* Oxford: Pergamon, 1985, 149-161.

Groot, P.J.M. Validation of language tests. L. Palmer et B. Spolsky (réd.). *Papers on Language Testing 1967-1974.* Washington, DC: TESOL, 137-143.

Grotjahn, R. On the methodological basis of introspective methods. In C. Faerch et G. Kasper (réd.), *Introspection in Second Language Research.* Clevedon: Multilingual Matters, 1987. 54-81.

Gulliksen, H. *Theory of Mental Tests.* New York: Wiley, 1950.

Guttman, L. A basis for analysing test-retest reliability. *Psychometrika*, 1945, *10*. 255-282.

Hambleton, R.K. et Cook, L.L. Latent trait models and their use in the analysis of educational test data. *Journal of Educational Statistics*, 1977, *14*. 75-96.

Hambleton, R.K. et Rovinelli, R.J. Assessing the dimensionality of a set of test items. *Applied Psychological Measurement.* 1986, *10*. 287-302.

Hambleton, R.K. et Swaminathan, H. *Item Response Theory: Principles and Application.* Boston: KluwerNijhoff, 1985.

Hambleton, R.K. et Traub, R.E. The effect of item order on test performance and stress. *Journal of Experimental Education.* 1974, *24*, 273-281.

Hanania, E. et Shikhani, M. Interrelation ships among three tests of language proficiency: Standardized ESL, Cloze, and writing. *TESOL Quarterly*, 1986, *20*, 97-109.

Harley, B., King, M.L. et Burtis, J. Perspective on lexical proficiency in a second language. In B. Harley et al. (réd.), *The Development of Bilingual Proficiency: Final Report* (vol. 1). Toronto: OISE, Modern Language Center, avril 1987.

Harris, D.P. *Testing English as a Second Language*. New York: McGraw-Hill, 1969.

_____ Report on an experimental group-administered memory span-test. *TESOL Quarterly*, 1970, *4*, 203-213.

Harrison, A. Communicative testing: Jam tomorrow? In A. Hughes et D. Porter (réd.), *Current Developments in Language Testing*. Londres: Academic Press, 1983, 77-85. (a)

_____ *A Language Testing Handbook*. Londres: Macmillan, 1983 (b)

Harrison, D.A. Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics*, 1986, *11*, 91-115.

Hattie, J.A. *Decision Criteria for Determining Unidimensionality*. Thèse de doctorat non-publiée, University of Toronto, OISE, 1981.

_____ Methodology Review: Assessing Unidimen sionality of tests and items. *Applied Psychological Measurement*, 1985, *9*, 139-164.

Henning, G. Advantages of latent trait measurement in language testing. *Language Learning*, 1984, *1*, 123-133.

_____ Item banking via DBase II: The UCLA ESL proficiency examination experience. In C.W. Stansfield (réd.), *Technology and Language Testing*. Washington, DC: TESOL, 1986, 71-77.

_____ *A Guide to Language Testing: Development, Evaluation, Research*. Cambridge, MA: Newbury House, 1987.

_____ The influence of test and sample dimensionality on latent trait person ability and item difficulty calibration. *Language Testing*, 1988, *5*, 83-99.

Henning, G., Hudson T. et Turner J. Item response theory and the assumption of unidimensionality for language tests. *Language Testing*, 1985, *2*, 141-154.

Henrysson, S. Gathering, analysing, and using data on test items. In R.L. Thorndike (réd.), *Educational Measurement* (2ème éd.). Washington, DC: American Council on Education, 1971, 130-159.

Higgins, J. et Johns, T. *Computers in Language Learning*. Londres: Addison-Wesley, 1984.

Hicks, M.M. Computerized multilevel ESL testing, a rapid screening methodology. In C.W. Stansfield (réd.). *Technology and Language Testing*. Washington, DC: TESOL, 1986, 79-90.

Hills, J.R. Use of measurement in selection and placement. In R.L. Thorndike (réd.). *Educational Measurement* (2ème éd.). Washington, DC: American Council on Education, 1971, 680-732.

Hinofotis, F.B. Cloze as an alternative method of ESL placement and proficiency testing. In J.W. Jr Oller et K. Perkins (réd.), *Research in Language Testing*. Rowley, MA: Newbury House, 1980, 121-128.

Hinofotis, F.B., Bailey, K.M. et Stern, S.L. Assessing the oral proficiency of prospective foreign teaching assistants: Instrument development. In A.S. Palmer, P.J.M. Groot et G.A. Trosper (réd.), *The Construct Validation of Tests of Communicative Competence*. Washington, DC: TESOL, 1981, 106-126.

Holmes, G et Kidd, M.E. Second-language learning and computers. *Revue canadienne des langues vivantes*, 1982, *38*, 503-516.

Howard, F. Testing communicative testing proficiency in French as a second language. *Revue canadienne des langues vivantes*, 1980, *36*, 272-283.

Hulin, C.I., Drasgow, F. et Parson, C.K. *Item Response Theory: Application to Psychological Measurement*. Homewood, IL: Dow Jones-Irwin, 1983.

Hulstijn, J.H. Second language proficiency: An interactive approach. In K. Hyntenstan et M. Pienemann (réd.), *Modelling and Assessing Second Language Acquisition*. Clevedon: Multilingual Matters, 1985, 373-380.

Ilyin, D. Structure placement test for adults in English-second-language programs in California. *TESOL Quarterly*, 1970, *4*, 129-136.

Ingram, D.E. Assessing proficiency: An overview on some aspects of testing. In K. Hyntenstan et M. Pienemann (réd.), *Modelling and Assessing Second Language Acquisition*. Clevedon: Multilingual Matters, 1985, 215-275.

_____ *The International English Language Testing System (IELTS): Its Nature and Development*. Communication au RELC Regional Seminar, Singapour, avril 1990.

Ingram, E. The psycholinguistic basis. In B. Spolsky (réd.), *Advances in Language Testing Series: 2 — Approaches to Language Testing*. Arlington, VA: Center for Applied Linguistics, 1978, 1-14.

Jafarpur, A. The short-context technique: an alternative for testing reading comprehension. *Language Testing*, 1987, *4*, 195-220.

Jarmasz, J.W. Pour une approche éclectique à l'évaluation de l'apprentissage linguistique de l'adulte. *Médium*, 1983, *6*, 4:85-90.

Jasmin-Demers L. Des tests pragmatiques et multifonctionnels pour une évaluation objective de la compétence à communiquer. *Médium*, 1983, *8*, 1:93-102.

Jensen, A.R. *Bias in Mental Tests*. New York: The Free Press, 1980.

Jochems, W. et Montens, F. The multiple-choice Cloze test as a general language proficiency test. *I.T.L.*, 1988, 81/82, 139-159.

Johnson, D.F. et Mihal, W.L. Performance on Blacks and Whites in computerized versus manual testing environment. *American Psychologist*, 1973, *28*, 694-699.

Johnson, K. Communicative approaches and communicative processes. In C.V. Brumfit et K. Johnson (réd.), *The Communicative Approach to Language Teaching*. Oxford: Oxford University Press, 1979, 192-205.

Jones, R.L. The FSI oral interview. In B. Spolsky (réd.), *Advance in Language Testing Series: 1 — Some Majors Tests*. Arlington, VA: Center for Applied Linguistics, 1978, 104-115.

_____ Second language performance testing: an overview. In P. Hauptman, R. Leblanc et M. Wesche (réd.), *L'évaluation de la performance en langue seconde*. Ottawa, ON: Éditions de l'Université d'Ottawa, 1985, 15-24.

Jonz, J. Improving on the basic egg: The M-C Cloze. *Language Learning*, 1976, *26*, 255-265.

Jöreskog, K.G. Statistical analysis of sets of congeneric tests. *Psychometrika*, 1971, *36*, 109-133.

Jöreskog, K.G. et Sörbom, D. *LISREL: User's Guide* (2ème éd.). Uppsala, Suède: National Educational Resources, 1983.

Kaya-Carton, E. et Carton, A.S. Multidimensionality of foreign language reading proficiency: Preliminary considerations in assessment. *Foreign Language Annals*, 1986, *18*, 95-102.

Keating, N. The Summer Language Bursary Program: A Canadian success story. *Revue canadienne des langues vivantes*, 1989, *45*, 457-463.

Kingsbury, G.G et Weiss, D.J. *An Adaptive Testing Strategy for Mastery Decisions* (Research Report 79-5). Minneapolis, MN: University of Minnesota, Department of Psychology, 1979.

_____ *A Comparison of Adaptive, Sequential and Conventional Testing Strategies for Mastery Decisions* (Research Report 80-4). Minneapolis, MN: University of Minnesota, Department of Psychology, 1980. (a)

_____ *An Alternate-Forms Reliability and Concurrent Validity Comparison of Bayesian Adaptive and Conventional Ability Tests* (Research Report 80-5). Minneapolis, MN: University of Minnesota, Department of Psychology, 1980. (b)

_____ *A Validity Comparison of Adaptive and Conventional Strategies for Mastery Testing* (Research Report 81-3. Minneapolis, MN: University of Minnesota, Department of Psychology, 1981.

_____ Adaptive Mastery Testing and Sequential Mastery Testing. In D.J. Weiss (réd.). *New Horizons in Testing*. New York: Academic Press, 1983, 257-283.

Klein-Braley, C. et Raatz, U. A survey of research on the C-test. *Language Testing*, 1984, *1*, 134-146.

Koch, B.R. et Patience, W.M. Students' attitudes toward tailored testing. In D.J. Weiss (réd.), *Proceeding of the 1977 Conference on Computerized Adaptive Testing*. Minneapolis, MN: University of Minnesota, 1978, 116-127.

Krashen, S. Is the "natural order" an artifact of the Bilingual Syntax Measure? *Language Learning*, 1978, *28*, 187-191.

_____ *Second Language Acquisition and Second Language Learning*. Oxford: Pergamon, 1981.

_____ *Principles and Practice in Second Language Learning*. Oxford: Pergamon, 1983.

Kreitzberg, C.B., Stocking, M.L. et Swanson, L. Computerized adaptive testing: Principles and directions. *Computers and Education*. 1978, *2*, 319-329.

Krzanowski, W.J. et Woods, A.J. Statistical aspects of reliability in language testing. *Language Testing*, 1984, *1*, 1-20.

Labelle, F. Des défis pour la linguistique appliquée. *Bulletin de l'ACLA*, 1986, *8*, 1:23-31.

Lado, R. *Language Testing: The Construction and Use of Foreign Language Tests*. Londres: Longman, 1961.

Lantolf, J.P. et Frawley, W. Proficiency: Understanding the construct. *Studies in Second Language Acquisition*, 1988, *10*, 181-196.

Lapkin, S. Pedagogical implications of direct second language testing: A Canadian example. In K. Hyntenstan et M. Pienemann (réd.), *Modelling and Assessing Second Language Acquisition*. Clevedon: Multilingual Matters, 1985, 333-347.

Lapkin, S. et al. Annotated list of French tests. *Revue canadienne des langues vivantes*, 1984, *41*, 93-109.

Larson, J.W. Computerized adaptive language testing: A Spanish placement exam. In K.M. Bailey, T.L. Dale et R.T. Clifford (réd.), *Language Testing Research*. Monterey, CA: Defence Language Institute, 1-23.

Larson, J.W. et Madsen, H.S. Computerized adaptive language testing: Moving beyond computer-assisted testing. *CALICO Journal*. Mars 1985, 32-43.

Laurier, M. Comparison of three different methods of assessing second-language proficiency. *York Working Papers in Second-Language Teaching*, 1986, *1*, 3-14.

Leblanc, R. De la mesure en toute chose. *Langue et société*, 1983, *9*, 10-13.

_____ Pour un classement sans douleur. *Bulletin de l'AQEFLS*, 1985, *7*. 1/2:7-15.

_____ L'évaluation en langue seconde dans un contexte communicatif. In R. Leblanc et al. (réd.), *L'enseignement des langues secondes aux adultes: Recherches et pratiques*. Ottawa, ON: Presses de l'Université d'Ottawa, 1989, 55-74.

Lee, Y.P Investigating the validity of the Cloze score. In Y.P. Lee (réd.), *New Directions in Language Testing*. Oxford: Pergamon, 1985, 137-147.

Levine, M.V. et Drasgow, F. Appropriate measurement: Validating studies and variables. In D.J. Weiss (réd.), *New Horizons in Testing*. New York. Academic Press, 1983, 109-131.

Lewkowicz, J.A. et Moon, J. Evaluation: A way of involving the learner. In J.C. Alderson (réd.), *Evaluation*. Oxford: Pergamon, 1985, 45-80.

Lin..., R.T., Rock, D.R. et Cleary, T.A. The development and evaluation of several programmed testing methods. *Educational and Psychological Measurement*, 196[illegible] *29*, 129-146.

Linn, R.L. et Werts, C.E. Covariance structures and their analysis. *New Directions for Testing and Measurements*, 1979, *4*, 53-73.

Lord, F.M. The relation of test scores to the trait underlying the test. *Educational and Psychological Measurement*, 1953, *13*, 517-548.

_____ Some test theory for tailored testing. In W.H. Holtzman (réd.), *Computer-Assisted Instruction, Testing and Guidance*. New York: Harper & Row, 1970, 139-183.

_____ A theoretical study of the measurement effectiveness of flexilevel tests. *Educational and Psychological Measurement*, 1971, *31*, 805-813.

_____ A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, 1977, *1*, 95-100. (a)

_____ Some item analysis and test theory for a system of computer-assisted test construction for individualized instruction. *Applied Psychological Measurement*, 1977, *1*, 447-455. (b)

_____ Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 1977, *14*, 117-138.

_____ *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Erlbaum, 1980.

_____ Small N Justifies Rasch Model. In D.J. Weiss (réd.), *New Horizons in Testing*. New York: Academic Press, 1983, 51-61.

Lord, F.M. et Novick, M.R. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley, 1968.

Low, G. Validity and the problem of direct language proficiency test. In J.C. Alderson (réd.), *Evaluation*. Oxford: Pergamon, 1985, 151-168.

Lowe, P. Jr. "The" question. *Foreign Language Annals*, 1984, *17*, 381-387.

_____ The ILR proficiency scale as a synthetisizing research principle: The view from the mountain. In C.J. James (réd.), *Foreign Language Proficiency in the Classroom and Beyond*. Lincolwood, IN: National Textbook, 1985, 9-54.

Lowe, P. Jr et Clifford, R.T. Developing an indirect measure of overall oral proficiency. In J.R. Firth (réd.), *Measuring Spoken Language Proficiency*. Washington, DC: Georgetown University Press, 1980, 31-39.

Lowe, P. Jr, Janczewski, D.W. et al. *The Implication of CAT for AEI reading proficiency testing: The Dutch example*. Communication au symposium de CALICO, Colorado Springs, avril 1989.

Lussier-Charles, D. et Danan, M. (réd.), *Guide d'évaluation en classe: primaire, langues secondes* (version préliminaire). Québec: Gouvernement du Québec, Ministère de l'Éducation, février 1983.

Madsen, H.S. Determining the debilitative impact of test anxiety. *Language Learning*, 1982, *32*, 133-143.

_____ Evaluating a computer-adaptive ESL placement test. *CALICO Journal*. Décembre 1986, 41-50.

_____ *Computest: ESL Version 2.5*. Orem, UT: CALI Inc., 1989. (a)

_____ *Introduction to Language Adaptive Testing*. Atelier au symposium de CALICO, Colorado Springs, mars 1989. (b)

Madsen, H.S. et Larson, J.W. Computerized Rasch analysis of item bias in ESL test. In C.W. Stansfield (réd.), *Technology and Language Testing*. Washington, DC: TESOL, 1986, 47-67.

Magnan, S.S. From achievement toward proficiency through multi-sequence evaluation. In C.J. James (réd.), *Foreign Language Proficiency in the Classroom and Beyond*. Lincolnwood, IN: National Textbook, 1985, 117-146.

Manning, W.H. Using technology to assess second-language proficiency through Cloze-elide tests. In C.W. Stansfield (réd.), *Technology and Language Testing*. Washington, DC: TESOL, 1986, 149-165.

Martin, J.T., McBride, J.R. et Weiss, D.J. *Reliability and Validity of Adaptive and Conventional Tests in a Military Recruit Population* (Research Report 83-1). Minneapolis, MN: University of Minnesota, Department of Psychology, 1983.

Mattran, K.J. Native speakers' reactions to speakers of ESL: Implication for adult basic education oral English proficiency testing. *TESOL Quarterly*, 1977, *11*, 407-414.

McDonald, R.P. The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 1980, *34*, 100-117.

_____ *Factor Analysis and Related Methods*. Hillsdale, NJ: Laurence Erlbaum, 1985.

McLean, L.D. et Ragsdale, R.G. The Rasch model for achieve ment tests - Inappropriate in the past, inappropriate today, inappropriate tomorrow. *Canadian Journal of Education*, 1983, *8*, 71-76.

McNamara, T.F. *The Role of Item Response Theory in Language Test Validation*. Communication au RELC Reginal Seminar, Singapour, avril 1990.

Meara, P. et Buxton, B. An alternative to multiple-choice vocabulary tests. *Language Testing*, 1987, *4*, 142-151.

Millman, J. et Arter, J.A. Issues in item banking. *Journal of Educational Measurement*, 1984, *21*, 315-330.

Mislevy, R.L. et Bock R.D. *PC-BILOG*. Mooresville, IN: Scientific Software, 1986.

Mitchell, J.V. (réd.). *Tests In Print III*. Lincoln, NB.: University of Nebraska Press.

Monfils, G. Étude de la validité du MLAT comme prédicteur du succès aux tests de rendement des fonctionnaires fédéraux canadiens dans l'apprentissage du français, langue seconde. *Médium*, 1982, 7, 2:21-46.

_____ L'évaluation de l'apprentissage au sein du programme langue de travail — un bilan positif. *Médium*, 1982, 7, 2:103-109.

Morris, C.N. On the foundations of test equating. In P.W. Holland et D.R. Rubin (réd.), *Test Equating*. New York: Academic Press, 1982, 169-191.

Morrow, K. Communicative language testing: revolution or evolution. In C.J. Brumfit et K. Johnson (réd.), *The Communicative Approach to Language Teaching*. Londres: Oxford University Press, 1979, 143-157.

_____ Testing spoken language. In J.B. Heaton (réd.), *Language Testing*. Middlesex: Modern English Publications, 1982, 56-58.

Mothe, J.-C. Les tests de savoir-faire en langue seconde: l'expérience européenne. In P. Hauptman, R. Leblanc et M. Wesche (réd.), *L'évaluation de la performance en langue seconde*. Ottawa, ON: Éditions de l'Université d'Ottawa, 1985, 59-70.

Mussio, J.L. *A Modification to Lord's Model for Tailored Tests*. Thèse de doctorat non publiée, University of Toronto, OISE, 1973.

Nagle, S.J. et Sanders, S.L. Comprehension theory and second language pedagogy. *TESOL Quarterly*, 1986, *20*, 12-21.

Nelson, H., Lomax, R.G. et Perlman, R. A structural equation model of second language acquisition for adult learners. *The Journal of Experimental Education*, 1984, *53*, 29-39.

Nelson, L.R. *Guide to LERTAP Use and Interpretation*. Dunedin, Nouvelle-Zélande: University of Otego, Education Department, 1970.

Nevo, B. Face validity revisited. *Journal of Educational Statistics*, 1985, *22*, 287-293.

Newsham, G.S. Communicative testing and classroom teaching. *Revue canadienne des langues vivantes*, 1989, *45*, 338-344.

Ng, E.K.L. et Ollivier, W. Computer-assisted language learning: An investigation on some design and implememtation issues. *System*, 1987, *15*, 1-17.

Nitko, A.J. et Hsu, T.-C. A comprehensive micro-computer system for classroom testing. *Journal of Educational Statistics*, 1984, *21*, 377-390.

Nie, N. et al. *SPSSX. User's Guide*. McGraw-Hill/SPSS Inc., 1983.

Norusis, M.J. et al. *SPSS/PC+ V2.0*. Spss Inc., 1988.

Oller, J.W. Jr. Assessing competence in ESL: Reading. *TESOL Quarterly*, 1972, 6, 26-36.

_____ Cloze tests of second language proficiency and what they measure. *Language Learning*, 1973, 23, 105-118.

_____ Pragmatics and language testing. In B. Spolsky (réd.), *Advances in Language Testing Series: 2 — Approaches to Language Testing*. Arlington, VA: Center for Applied Linguistics, 1978, 39-57.

_____ *Language Tests at School*. Londres: Longman, 1979.

_____ Is there a global factor of language proficiency? In J.A.S. Read (réd.), *Directions in Language Testing*. Singapour: Singapore University Press, 1981, 3-40.

_____ Evidence for a general proficiency fac tor: an expectancy grammar. In J.W. Jr Oller (réd.), *Issues in Language Testing Research*. Rowley, MA: Newbury House, 1983, 3-10.

Oller, J.W. Jr et Hinofotis, F.B. Two mutually exclusive hypotheses about second language ability: indivisible or partially divisible competence. In J.W. Jr Oller et K. Perkins (réd.), *Research in Language Testing*. Rowley, MA: Newbury House, 1980, 13-23.

Oller, J.W. Jr et Inal, N. A Cloze test of English preposition. *TESOL Quarterly*, 1971, 5, 315-326.

Oller, J.W. Jr et Perkins, K. (réd.). *Research in Language Testing*. Rowley, MA: Newbury House, 1980.

Oller, J.W. Jr et Streiff, V. Dictation: A test of grammar based expectancies. In R.L. Jones et B. Spolsky (réd.), *Testing Language Proficiency*. Arlington, VA: Center for Applied Linguistics, 1975, 71-88.

Olsen, S. Foreign language departments and computer-assisted instruction: A survey. *Modern Language Journal*, 1980, 64, 341-349.

Omaggio, A.C. Methodology in transition: The new focus on proficiency. *Modern Language Journal*, 1983, 67, 330-341.

_____ *Proficiency-Oriented Classroom Testing*. Washington, DC: Center for Applied Linguistics/ERIC, 1983.

O'Malley, J.M. et al. Learning strategies applications with students of English as a second language. *TESOL Quarterly*, 1985, 19, 557-584.

Oskarsson, M. Subjective and objective assessment of foreign language performance. In J.A.S. Read (réd.), *Directions in Language Testing*. Singapour: Singapore University Press, 1981, 225-239.

Painchaud, G. et Leblanc, R. L'auto-évaluation en contexte scolaire. In C. Germain (réd.), *Études de linguistique appliquée 56*. Paris: Didier, 1984, 88-98.

Palmer, A.S. Compartmentalized and integrated control. In J.W.Jr Oller (réd.), *Issues in Language Testing Research*. Rowley, MA: Newbury House, 1983, 323-332.

Perkins, K. et Miller, L.D. Comparative analysis of English as a second language reading comprehension data. *Language Testing*, 1984, *1*, 21-32.

Petersen, C.R. et Cartier, F.A. Some theoretical problems and practical solutions in proficiency test validity. In R.L. Jones et B. Spolsky (réd.). *Testing Language Proficiency*. Arlington, VA: Center for Applied Linguistics, 1975, 105-118.

Pine, S.M. Reduction of test bias by adaptive testing. In D.J. Weiss (réd.), *Proceeding of the 1977 Conference on Computerized Adaptive Testing*. Minneapolis, MN: University of Minnesota, 1978, 128-142.

Pine, S.M. et Weiss, D.J. *A Comparison of the Fairness of Adaptive and Conventional Testing Strategies* (Research Report 78-1). Minneapolis, MN: University of Minnesota, Department of Psychology, 1978.

Pine, S.M. et al. *Effects of Computerized Adaptive Testing on Black and White students*. (Research Report 79-2). Minneapolis, MN: University of Minnesota, Department of Psychology, 1979.

Porter, D. The effect of quantity of context on the ability to make linguistic predictions. In A. Hughes et D. Porter (réd.), *Current Developments in Language Testing*. Londres: Academic Press, 1983, 63-74.

Potts, P.J. The role of evaluation in a communicative curriculum, and some consequences for material design. In J.C. Alderson (réd.), *Evaluation*. Oxford: Pergamon, 1985, 19-44.

Prestwood, J.S. Effects of knowledge of results and varying proportion correct on ability test performance and psychological variables. In D.J. Weiss (réd.), *Proceeding of the 1977 Conference on Computerized Adaptive Testing*. Minneapolis, MN: University of Minnesota, 1978, 106-115.

Prestwood, J.S. et Weiss, D.J. *Accuracy of Perceived Test-Item Difficulties* (Research Report 77-3). Minneapolis, MN: University of Minnesota, Department of Psychology, 1977.

Raatz, U. Better theory for better tests? *Language Testing*, 1985, 2, 60-75.

Raffaldini, T. The use of situation tests as measure of communicative ability. *Studies in Second Language Acquisition*, 1988, *10*, 197-215.

Ramirez, A.G. Pupil characteristics and communicative language measures. In C. Rivera (réd.), *Communicative Competence Approaches to Language Proficiency Assessment: Research and Applications*. Clevedon: Multilingual Matters, 1984, 82-106.

Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhague. Danish Institute for Educational Research, 1960.

Raschio, R.A. Communicative uses of the computer: Ideas and directions. *Foreign Language Annals*, 1986, *19*, 507-514.

Reckase, M.D. A procedure for decision making using tailored testing. In D.J. Weiss (réd.), *New Horizons in Testing*. New York: Academic Press, 1983, 237-255.

Ree, M.J. The effects of item calibration sample size and item pool on adaptive testing. *Applied Psychological Measurement*, 1981, 5, 11-19.

Ricciardi, J.S. Second language tests for specific purposes: How specific? *Médium*, 1981, *6*, 4:143-160.

Rivera, C. et Simich, C. Language proficiency assessment: Research finding and their applications. In S. Seidner (réd.), *Issues of Language Assessment: Foundation and Research*. Evanston, IL: Illinois State Board of Education, 1982, 37-41.

Roe, P.J. Une réévaluation de l'évaluation ou «Le coucou dans le nid». *Le français dans le monde*, 1981, *165*, 33-47.

Rushinek, A., Rushinek, S.F. et Stutz, J. Relationship of computer users' performance to their attitudes toward interactive software. *Journal of Educational Technology Systems*, 1985, *13*, 255-264.

Sachar, J.D. et Fletcher, J.D. Administrating paper-and-pencil test by computer, or the medium is not always the message. In D.J. Weiss (réd.), *Proceeding of the 1977 Conference on Computerized Adaptive Testing*. Minneapolis, MN: University of Minnesota, 1978, 403-422.

Samejima, F. A use of the information function in tailored testing. *Applied Psychological Measurement*, 1977, *1*, 233-247.

_____ The application of graded response models: the promise of the future. In D.J. Weiss (réd.), *Proceeding of the 1977 Conference on Computerized Adaptive Testing*. Minneapolis, MN: University of Minnesota, 1978, 28-37.

Saracho, O.N. Teaching second-language literacy with computers. In D. Hainline (réd.), *New Developments in Computer-Assisted Language Learning*. Londres: Croom Helm, 1987, 53-68.

Savard, J.-G. *Bibliographie analytique des tests de langue*. Québec: C.I.R.B., Presses universitaires de l'Université Laval, 1969.

Savignon, S.J. *Communicative Competence: Theory and Classroom Practice*. Reading, MA: Addison-Wesley, 1983.

_____ Evaluation of communicative competence: The ACTFL proficiency guidelines. *Modern Language Journal*, 1985, 69, 129-134.

Saville-Troike, M. What ***really*** matters in second language learning for academic achievement? *TESOL Quarterly*, 1984, *18*, 199-219.

Scholz, G. et al. Is language ability divisible or unitary? A factor analysis of 22 English language proficiency tests. In J.W.Jr Oller et K. Perkins (réd.), *Research in Language Testing*. Rowley, MA: Newbury House, 1980, 24-33.

Scott, M. Students' affective reaction to oral language test. *Language Testing*, 1986, *3*, 99-118.

Seaton, I. The English Language Testing Service (ELTS): Two issues in the design of the new non-academic module. In A. Hughes et D. Porter (réd.), *Current Developments in Language Testing*. Londres: Academic Press, 1983, 129-140.

Séguin, S.S. *An Exploratory Study of the Efficiency of the Flexilevel Testing Procedure*. Thèse de doctorat non publiée, University of Toronto, OISE, 1976.

Séguin, S.S. et Auger, R. Une introduction à la théorie des réponses aux items. *Mesure et évaluation en éducation*, 1986, *9*, 1:7-44.

Séguin-Duquette, L. La théorie du trait latent. *Médium*, 1982, *7*, 63-69.

Shohamy, E. Affective considerations in language testing. *Modern Language Journal*, 1982, 66, 13-17.

_____ Does the testing method make the difference? The case of reading comprehension. *Language Testing*, 1984, *1*, 147-170.

Shohamy, E. et Reves, T. Authentic language tests: Where from and where to? *Language Testing*, 1985, *2*, 48-59.

Spolsky, B. Language testing—The problem of validation. *TESOL Quarterly*, 1968, *2*, 88-94.

_____ Language testing: Art or science? In G. Nickel (réd.), *Actes du 4ème congrès annuel de linguistique appliquée 3*. Stuttgard: Hochschul-Verlag, 1976, 215-235.

_____ What does it mean to know how to use a lan guage? An essay on the theoretical basis of language testing. *Language Testing*, 1985, 2, 180-191.

Stansfield, C.W. Reliability of the secondary level English proficiency tests. *System*, 1982, 12, 1-11.

Stevensen, D.K. Pop validity and performance testing. In Y.P. Lee (réd.), *New Directions in Language Testing*. Oxford: Pergamon, 1985, 11-118.

Streiff, V. The role of language in educational testing. In J.W. Jr Oller (réd.), *Issues in Language Testing Research*. Rowley, MA: Newbury House, 1983, 343-350.

Swain, M. Teaching and testing communicatively. *TESL Talk*, 1984, 15, 7-18. (a)

_____ Large-scale communicative language testing: A case study. In S.J. Savignon et M.S. Berns (réd.), *Initiatives in Communicative Language Testing*. Reading, MA: Addison-Wesley, 1984, 186-201.

Swain, M, Dumas, G. et Naiman, N. Alternatives to spontaneous speech: Elicited translation and imitation as indicators of second language competence. *Working Papers in Bilingualism*, 1974, 3, 68-79.

Tatsuoka, K.K. et Tatsuoka, M.M. Detection of aberrant responses patterns and their effect on dimensionality. *Journal of Educational Statistics*, 1982, 7, 215-231.

Theunissen, T.J.J.M. Test banking and test design. *Language Testing*, 1987, 4, 1-8.

Thissen, D. *MultiLOG: User's Guide* (version 5). Mooresville, IN: Scientific Software, 1986.

Thissen, D. et Wainer, H. Some standard errors in item response theory. *Psychometrika*, 1982, 47, 397-412.

Thrush, J. et Thrush, R.S. Microcomputers in foreign language instruction. *Modern Language Journal*, 1984, 68, 21-26.

Traub, R.E. A priori considerations in choosing an item response model. In R.K. Hambleton (réd.), *Applications of Item Response Theory*. Vancouver, BC: Vancouver Educational Research Institute of British Columbia, 1983, 57-70.

Traub, R.E. et Lam, Y.R. Latent structure and item sampling models for testing. *Annual Review of Psychology*, 1985, 36, 19-48.

Tung, P. Computerized adaptive testing: Implications for language test developers. In C.W. Stansfield (réd.), *Technology and Language Testing*. Washington, DC: TESOL, 1986, 11-28.

Tyler, R.W. Using tests in grouping students for instruction. In R.W. Tyler et R.M. Wolf (réd.), *Crucial Issues in Testing*. Berkeley, CA: McCutchan, 1974, 65-70.

Upshur, J.A. et Homburg, T.J. Some relations among language tests at successive ability levels. In J.W. Jr Oller (réd.), *Issues in Language Testing Research*. Rowley, MA: Newbury House, 1983, 188-202.

Urry, V.W. Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement*. 1977, *14*, 181-196.

Vale, C.D. Linking item parameters onto a common scale. *Applied Psychological Measurement*, 1986, *10*, 333-344.

Vale, C.D. et Gialluca, K.A. Evaluation of the efficiency of item calibration. *Applied Psychological Measurement*, 1988, *12*, 53-67.

Vale, C.D. et Weiss, D.J. *A Simulation Study of Stradaptive Ability Testing* (Research report 75-6). Minneapolis, MN: University of Minnesota, Department of Psychology, 1975.

Valette, R.M. *Modern Language Testing* (2ème éd.). New York: Hartcourt Brace Johanovitch, 1977.

Vetterli, C.F. *Efficacity of different Item Bias Detection Methods in Detecting Multidimensional Bias*. Thèse de maîtrise non publiee, University of Toronto, OISE, 1987.

Vollmer, H.J. Why are we interested in general language proficiency? In C. Klein-Bradley (réd.), *Practice and Problems in Language Testing*. Francfort: Verlag Peter D. Lang, 1981, 96-123.

_____ The structure of foreign language competence. In A. Hughes et D. Porter (réd.), *Current Developments in Language Testing*. Londres: Academic Press, 1983, 3-29.

Wainer, H. On item response theory and computerized adaptive tests. *The Journal of College Admissions*, 1983, *28*, 4:9-16.

Wald, A. *Sequential Analysis*. New York: Wiley, 1947.

Watts, C. *Interactive Video: What the Students Say*. Communication au symposium de CALICO, Colorado Springs, mars 1989.

Weiss, D.J. Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 1982, *6*, 473-492.

Weiss, D.J. et Betz, N.E. *Effects of Immediate Knowledge of Results and Adaptive Testing of Ability Test Performance* (Research report 76-3). Minneapolis, MN: University of Minnesota, Department of Psychology, 1976. (a)

_____ *Psychological Effects of Im mediate Knowledge of Results and Adaptive Testing* (Research report 76-3). Minneapolis, MN: University of Minnesota, Department of Psychology, 1976. (b)

Weiss, D.J. et Brown, J.M. Multi-content adaptive measurement of achievement. In D.J. Weiss (réd.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis, MN: University of Minnesota, 1978, 313-330.

Weiss, D.J. et Kingsbury, G.G. Applications of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 1984, *21*, 361-375.

Wesche, M.B. Communicative testing in a second language. *Revue canadienne des langues vivantes*, 1981, *37*, 551-571.

_____ Second language performance testing: the Ontario Test of ESL as an example. *Language Testing*, 1987, *4*, 28-47.

Whitney, D.R. Credit recommendation for oral proficiency. In J.R. Firth (réd.), *Measuring Spoken Language Proficiency*. Washington, DC: Georgetown University Press, 1980, 60-63.

Wilcox, R.R. Solving measurement problems with an answer-until-correct scoring procedure. *Applied Psychological Measurement*, 1981, *5*, 399-414.

Wilds, C.P. The oral interview test. In R.L. Jones et B. Spolsky (réd.), *Testing Language Proficiency*. Arlington, VA: Center for Applied Linguistics, 1975, 29-44.

Willmot, A. et Kam Chuan Aik. *The Ngee Ann-Oxford Computerized English language Test (CELPT)*. Communication au RELC Regional Seminar, Singapour, avril 1990.

Wingersky, M.S., Barton, M.A. et Lord, F.M. *LOGIST User's Guide*. Princeton, NJ: Educational Testing Service, 1982.

Woods, A. Principal components and factor analysis in the investigation of the structure of language proficiency. In A. Hughes et D. Porter (réd.), *Current Developments in Language Testing*. Londres: Academic Press, 43-52.

Wright, B.D. et Bell, S.R. Item banks: What, why, how? *Journal of Educational Measurement*, 1984, *21*, 331-345.

Wright, B.D., Bell, S.R. et Mead, R.J. *BICAL: Calibrating Items with the Rasch model* (Version 3). Chicago: The University of Chicago, Department of Education, juin 1979.

Wright, B.D. et Pachapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, *29*, 23-48.

Wright, B.D. et Stone, M.H. *Best Test Design*. Chicago: MESA: 1979.

Yen, W.M. Use of the three-parameter logistic model in the development of a standardized achievement test. In R.K. Hambleton (réd.), *Applications of Item Response Theory*. Vancouver, BC: Vancouver Educational Research Institute of British Columbia, 1983, 123-141.

_____ Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement*, 1984, *21*, 93-111.

Young, H.T. On using Foreign Service Institute tests and standards on campuses. In J.R. Firth (réd.), *Measuring Spoken Language Proficiency*. Washington, DC: Georgetown University Press, 1980, 64-69.

Zeidner, M. et Bensoussan, M. College student's attitude towards written versus oral tests of English as a foreign language. *Language Testing*, 1988, 5, 100-114.

Zettensten, A. Experiments in large-scale vocabulary testing. In Y.P. Lee (réd.), *New Directions in Language Testing*. Oxford: Pergamon, 1986, 69-73.

Michel Laurier *est professeur à la Faculté des sciences de l'éducation de l'Université de Montréal. Il donne des cours en mesure et évaluation dans le cadre de la formation des maîtres. Il a aussi enseigné le français comme langue seconde pendant une quinzaine d'années. Il détient une maîtrise en linguistique appliquée de l'Université d'Ottawa et un doctorat de l'Institut d'études pédagogiques de l'Ontario. Ses champs de recherche sont l'évaluation de la compétence langagière et le testing informatisé. Il s'intéresse aussi à la pédagogie en enseignement supérieur.*

Le présent ouvrage retrace les étapes de la mise au point d'un test de classement en français langue seconde et discute des difficultés et des avantages de l'informatisation de cet instrument.